

Margins of discrete Bayesian networks.

Robin J. Evans

January 12, 2015

Abstract

In this paper we provide a complete algebraic characterization of the model implied by a Bayesian network with latent variables when the observed variables are discrete. We show that it is algebraically equivalent to the so-called nested Markov model, meaning that the two are the same up to inequality constraints on the joint probabilities. The nested Markov model is therefore the best possible approximation to the latent variable model whilst avoiding inequalities, which are extremely complicated in general. Latent variable models also suffer from difficulties of unidentifiable parameters and non-regular asymptotics; in contrast the nested Markov model is fully identifiable, represents a curved exponential family of known dimension, and can easily be fitted using an explicit parameterization.

1 Introduction

Directed acyclic graph (DAG) models, also known as Bayesian networks, are widely used multivariate models in probabilistic reasoning, machine learning and causal inference (Bishop, 2007; Darwiche, 2009; Pearl, 2009). These models are defined by simple factorizations of the joint distribution, and in the case of discrete or jointly Gaussian random variables, are curved exponential families of known dimension. The inclusion of latent variables within Bayesian network models can greatly increase their flexibility, and also account for unobserved confounding. However, this flexibility comes at the cost of creating models that are very complex, and that are not easy to explicitly describe when considered as marginal models over the observed variables. Latent variable models generally do not have fully identifiable parameterizations, and contain singularities that lead to non-regular asymptotics (Drton, 2009a). In addition, using them may force a modeller to specify a parametric structure over the latent variables, introducing additional assumptions that are generally difficult to test and may be unreasonable.

If no parametric assumptions are made about the latent variables, and no assumption is made about its state-space, this leads to an implicitly defined *marginal model*. The marginal DAG model has the advantage of avoiding

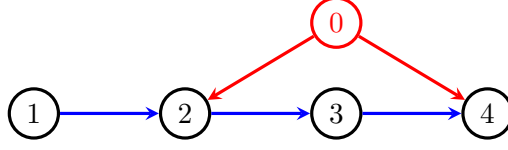


Figure 1: A directed acyclic graph on five vertices.

some of the assumptions made by a parametric latent variable model, but no explicit characterization of the model is available and nor is there any obvious method for fitting it to data.

Example 1.1. Consider the DAG on five vertices shown in Figure 1. The graph represents a multivariate model over five random variables X_0, X_1, X_2, X_3, X_4 with the restriction that the joint distribution factorizes as

$$p(x_0, x_1, x_2, x_3, x_4) = p(x_0) \cdot p(x_1) \cdot p(x_2 | x_0, x_1) \cdot p(x_3 | x_2) \cdot p(x_4 | x_0, x_3);$$

here, for example, $p(x_3 | x_2)$ represents the conditional density of X_3 given X_2 . If we treat X_0 as a latent variable, the *marginal model* over the remaining observed variables (X_1, X_2, X_3, X_4) is the collection of probability distributions that can be written in the form

$$\begin{aligned} p(x_1, x_2, x_3, x_4) \\ = \int_{\mathbf{x}_0} p(x_0) \cdot p(x_1) \cdot p(x_2 | x_0, x_1) \cdot p(x_3 | x_2) \cdot p(x_4 | x_0, x_3) dx_0. \end{aligned} \quad (1)$$

That is, any (X_1, X_2, X_3, X_4) -margin of a distribution which factorizes according to the DAG over all five variables, for any state-space or distribution of X_0 ¹.

From either of the displayed equations above we can deduce that $X_3 \perp\!\!\!\perp X_1 | X_2$, so this constraint holds in the marginal model. In other words, the conditional distribution $p(x_3 | x_1, x_2)$ does not depend upon x_1 . In addition this model satisfies the so-called *Verma constraint* of Verma and Pearl (1991), because the expression

$$\int_{\mathbf{x}_2} p(x_2 | x_1) \cdot p(x_4 | x_1, x_2, x_3) dx_2 \quad (2)$$

does not depend upon x_1 (see Example 3.2).

If the four observed variables are binary, the set of distributions satisfying the independence and Verma constraints is an 11-dimensional subset of the 15-dimensional probability simplex. It is not immediately clear whether or not this set is the same as the marginal model, since in principle there

¹In fact, without loss of generality we can assume X_0 is uniform on $(0, 1)$

might be other restrictions. In other words, can *any* distribution satisfying the constraints can be written in the form (1)? In this paper we will show that the constraints *are* sufficient to describe the model, up to inequalities². The marginal model is indeed 11-dimensional, and is algebraically defined by the equalities discussed above.

Existing approaches to this problem include the ancestral graph models of Richardson and Spirtes (2002) and the equivalent³ models on acyclic directed mixed graphs (ADMGs) of Richardson (2003). The Markov property considered by Richardson (2003) considers only conditional independence constraints and, in general, defines a strictly larger model than any latent variable model: we call this the *ordinary Markov model*. These models are the basis of the FCI algorithm for causal discovery in the presence of hidden variables (Spirtes et al., 2000).

The more refined *nested Markov property* (Shpitser et al., 2014) for ADMGs accounts for both conditional independences and Verma constraints. Though the nested model is smaller than the ordinary Markov model, it is known still to be strictly larger than the marginal model of interest because marginal models are subject to inequality constraints (Pearl, 1995; Evans, 2012).

1.1 Contribution

In this paper we show that marginal models with finite discrete observed variables are always algebraically fully described by the nested Markov property, in the sense that the Zariski closures of the marginal model and the nested model are the same.

A consequence of this is that a margin of a DAG model and its nested counterpart have the same dimension, and they differ only by inequality constraints. The situation is represented by Figure 2, which shows the marginal model lying strictly within the nested model, but the two sharing a tangent space at some point p_0 . This means that we have, for the first time, a full algebraic characterization of margins of Bayesian network models.

It also means that the nested model represents a sensible and practical approximation to the marginal model: inequality constraints are typically extremely complicated, so the nested model with its factorization criterion, separation criteria, and discrete parameterization, make it much easier to work with. The parameterization means that nested models can easily be fitted with existing algorithms (Evans and Richardson, 2010). In addition, the nested model is regular whenever the joint distribution is positive, so in a suitable sense it has better statistical properties than the marginal model.

²In fact, additional inequalities are present in this example, so strictly the answer to the question is ‘no’.

³The models are equivalent in the context we consider, but not if selection variables are present.

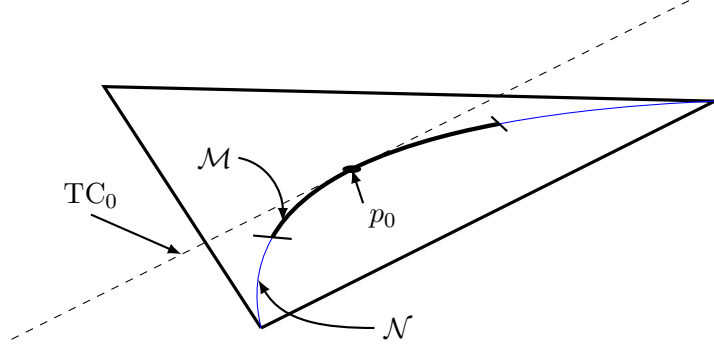


Figure 2: Diagrammatic representation of the marginal model (\mathcal{M}) sitting strictly within the nested model (\mathcal{N}), but sharing the same tangent cone TC_0 at the uniform distribution p_0 .

In principle causal discovery algorithms such as the FCI algorithm, which currently only make use of conditional independence constraints, could be extended to nested models. Our main result tells us that the nested model gives us close to maximum power to distinguish between different causal structures, without making additional assumptions.

We work with a class of hyper-graphs called mDAGs, with which we associate the marginal models of DAGs (Evans, 2014). The remainder of the paper is organized as follows: Section 2 introduces DAG models, their margins and mDAGs, and carefully defines the problem of interest. Section 3 describes the nested Markov property, and Section 4 gives an outline of the main result. Section 5 describes reductions which can be made to the state-space of the latent variables models without loss of generality, and Section 6 the main results of the paper. Finally in Section 7 we show that a large class of marginal models represent smooth manifolds, and provide some discussion.

2 Directed Graphical Models

We begin with some elementary graphical definitions.

Definition 2.1. A *directed graph*, $\mathcal{G}(V, \mathcal{E})$, consists of a finite set of vertices, V , and a collection of edges, \mathcal{E} , which are ordered pairs of distinct elements of V . If $(v, w) \in \mathcal{E}$ we denote this by $v \rightarrow w$, and say that v is a *parent* of w ; the set of parents of w is denoted $\text{pa}_{\mathcal{G}}(w)$. Similarly w is a *child* of v , and the child sets are denoted $\text{ch}_{\mathcal{G}}(v)$.

A directed graph is *acyclic* if there is no sequence of vertices $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_k \rightarrow v_1$ for $k > 1$. We call such a graph a *directed acyclic graph*, or DAG.

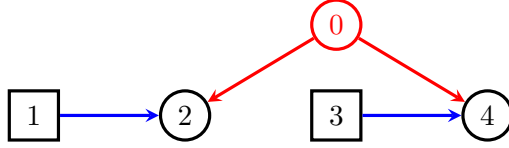


Figure 3: A directed acyclic graph on three random and two fixed vertices.

Graphs are best understood visually: an example of a DAG with five vertices and five edges is given in Figure 1.

We will require a very slight generalization of a DAG which introduces a second type of node.

Definition 2.2. A *conditional DAG* $\mathcal{G}(V, W, \mathcal{E})$ is a DAG with vertices $V \dot{\cup} W$ and edge set \mathcal{E} , with the restriction that no vertex in W may have any parents. The vertices in V are the *random vertices*, and W the *fixed vertices*; these two sets are disjoint.

If $W = \emptyset$, this reduces to the ordinary definition of a DAG. We denote fixed vertices with square nodes, and random ones with round nodes: see the example in Figure 3.

2.1 Graphical Models

A graphical model arises from the identification of a graph with a collection of multivariate probability distributions; see Lauritzen (1996) for an introduction. We associate each vertex $v \in V \cup W$ with a random variable X_v taking values in a finite state-space \mathfrak{X}_v . With a conditional DAG \mathcal{G} we associate some conditional probability measure P on $\mathfrak{X}_V \equiv \times_{v \in V} \mathfrak{X}_v$ given $\mathfrak{X}_W \equiv \times_{w \in W} \mathfrak{X}_w$; this distribution is subject to constraints determined by the structure of the graph.

Definition 2.3. Let P be a conditional probability distribution over \mathfrak{X}_V given \mathfrak{X}_W with conditional density p . We say that p obeys the *factorization criterion* with respect to a DAG \mathcal{G} if it factorizes into univariate conditional densities p_v , $v \in V$ as

$$p(x_V | x_W) = \prod_{v \in V} p(x_v | x_{\text{pa}(v)}), \quad x_{VW} \in \mathfrak{X}_{VW}. \quad (3)$$

The definition reduces to the familiar factorization criterion for DAGs if $W = \emptyset$. The extra generality will be useful for discussing Markov properties which involve factorization of the distribution into conditional pieces. The fixed vertices represent variables that have been conditioned upon; p satisfies (3) if and only if, after renormalization, it also satisfies the factorization criterion for the same DAG with all vertices random.

The definition of a Bayesian network can be extended to the case where no joint density exists by insisting that each random variable X_v can be written as a measurable function of $X_{\text{pa}(v)}$ and an independent noise variable; we call this the *structural equation property*. If the density exists the two criteria are equivalent, and since we work with discrete variables this condition is always satisfied. Although the factorization property is often simpler to work with for practical purposes such as modelling and fitting, the structural equation property is useful in proofs. The well-known global Markov property based on d-separation is also equivalent to the structural equation property (Pearl, 2009).

Example 2.4. A distribution P with density p obeys the factorization criterion for the graph in Figure 1 if the density has the form

$$p(x_0, x_1, x_2, x_3, x_4) = p(x_0) \cdot p(x_1) \cdot p(x_2 | x_0, x_1) \cdot p(x_3 | x_2) \cdot p(x_4 | x_0, x_3).$$

Such distributions are precisely those which satisfy the conditional independences

$$X_1 \perp\!\!\!\perp X_0, \quad X_3 \perp\!\!\!\perp X_0, X_1 | X_2, \quad X_4 \perp\!\!\!\perp X_1, X_2 | X_0, X_3.$$

Example 2.5. A conditional density obeys the factorization criterion for the conditional DAG in Figure 3 if it can be written as

$$p(x_0, x_2, x_4 | x_1, x_3) = p(x_0) \cdot p(x_2 | x_0, x_1) \cdot p(x_4 | x_0, x_3).$$

Latent Variables

We now introduce the possibility that some of the random vertices are unobserved, or *latent*. This leads to a model defined by integrating a factorization of the form above over the latent variables to obtain a marginal distribution.

Definition 2.6. Let \mathcal{G} be a conditional DAG with fixed vertices $V \cup U$, and random vertices W . A conditional density $p(x_V | x_W)$ is in the *V-marginal DAG model* for \mathcal{G} if there exists a density $q(x_V, x_U | x_W)$ such that q factorizes according to \mathcal{G} , and

$$p(x_V | x_W) = \int_{\mathbf{x}_U} q(x_V, x_U | x_W) dx_U.$$

That is, the margin of q over V is p .

Note that in principle this definition can be altered so as not to require the existence of a density; however, since the observed variables are all discrete, and the latent variables will be assumed to be independent of each other, assuming the existence of a density does incur any loss of generality.

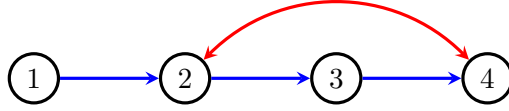


Figure 4: An mDAG representing the DAG in Figure 1, with the vertex 0 treated as unobserved.

2.2 mDAGs

We will represent the collection of marginal models defined by DAGs using a larger class of graphical models called mDAGs (‘marginal DAGs’). These avoid dealing with latent variables directly, instead introducing additional edges to represent them. For example, the DAG in Figure 1, with the vertex 0 treated as a latent variable, is represented by the mDAG in Figure 4.

Definition 2.7. An *mDAG*, $\mathcal{G}(V, W, \mathcal{E}, \mathcal{B})$, is hyper-graph consisting of a conditional DAG with random vertices V , fixed vertices W and directed edge set \mathcal{E} , together with a collection of *bidirected* hyper-edges \mathcal{B} : the elements of \mathcal{B} are inclusion maximal subsets of V , each of size at least two.

The mDAG was introduced by Evans (2014), without the additional generality of fixed vertices. This aspect changes very little to the theory of these graphs, but is necessary for understanding the nested Markov model. As with conditional DAGs, when representing mDAGs graphically the fixed vertices are drawn as square nodes and random vertices as circles. Bidirected edges are drawn in red, as in Figure 5(a); in this case $W = \{6\}$ and $\mathcal{B} = \{\{1, 2\}, \{2, 3, 4\}, \{3, 4, 5\}\}$.

With each mDAG, \mathcal{G} , we can associate a conditional DAG $\bar{\mathcal{G}}$ by replacing each bidirected edge $B \in \mathcal{B}$ with a new random vertex u , such that the children of u are precisely the vertices in B . The new vertex u becomes the ‘unobserved’ variable represented by the bidirected edge B . We call $\bar{\mathcal{G}}$ the *canonical DAG* associated with \mathcal{G} . The mDAG in Figure 5(a) is thus associated with the canonical DAG in Figure 5(b).

Our interest in mDAGs lies in their representation of the margin of the associated canonical DAG, and so we define our model in this spirit; see Evans (2014).

Definition 2.8. Let \mathcal{G} be an mDAG with vertices $V \cup W$. The *marginal model* for \mathcal{G} is the V -marginal model for $\bar{\mathcal{G}}$, the canonical DAG associated with \mathcal{G} . Denote the collection of such distributions by $\mathcal{M}(\mathcal{G})$.

In other words, the marginal model is the collection of distributions which could be constructed as the margin of a Bayesian network with latent variables replacing the bidirected edges. Any latent variable model (i.e. possi-

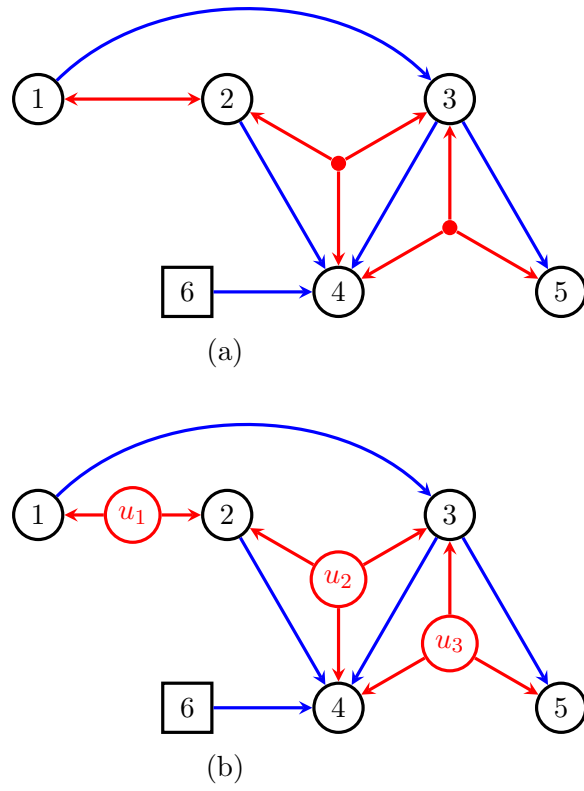


Figure 5: (a) An mDAG, \mathcal{G} , and (b) a DAG with hidden variables, $\bar{\mathcal{G}}$, representing the same model (the canonical DAG).

bly with parametric or other distributional assumptions on the latent variables) lies within the marginal model. If \mathcal{G} is a (conditional) DAG then the marginal model is just the usual model defined by the factorization.

From the definitions above, the set of marginal DAG models that can be represented by marginal models for mDAGs appears to be restricted to cases where the latent variables have no parents. In fact this does not cause any loss of generality (see Evans, 2014).

Just as distinct DAGs may be Markov equivalent, distinct mDAGs may give rise to the same marginal model: for example the graphs $1 \leftarrow 2 \leftarrow 3$ and $1 \leftarrow 2 \rightarrow 3$ and $1 \leftrightarrow 2 \rightarrow 3$ give rise to the same model. Some partial equivalence results for mDAGs are presented in Evans (2014).

Definition 2.9. A collection of (random) vertices $C \subseteq V$ in an mDAG \mathcal{G} is *bidirected-connected* if for any distinct $v, w \in C$, there is a sequence of vertices $v = v_0, v_1, \dots, v_k = w$ all in C such that, for each $i = 1, \dots, k$, the pair $\{v_{i-1}, v_i\}$ is contained in some bidirected edge in \mathcal{G} .

A *district* of an mDAG is an inclusion maximal bidirected-connected set of vertices.

More informally, a district is a maximal set of vertices joined by the red edges in an mDAG. It is easy to see from the definition that districts form a partition of the random vertices in an mDAG. The mDAG in Figure 4, for example, contains three districts, $\{1\}$, $\{3\}$ and $\{2, 4\}$. Districts inspire a useful reduction of mDAGs, via the following special subgraph.

Definition 2.10. Let \mathcal{G} be an mDAG containing vertices $D \subseteq V$. Then $\mathcal{G}[D]$ is the subgraph of \mathcal{G} with

- (i) random vertices D and fixed vertices $\text{pa}_{\mathcal{G}}(D) \setminus D$;
- (ii) those directed edges $w \rightarrow v$ such that $w \in D \cup \text{pa}_{\mathcal{G}}(D)$ and $v \in D$;
- (iii) the bidirected edges $\{B \cap D : B \in \mathcal{B}(\mathcal{G}) \text{ and } |B \cap D| \geq 2\}$.

$\mathcal{G}[D]$ is therefore the subgraph induced over D , together with parents of D and edges directed towards D . Any edges between parent vertices (whether directed or bidirected) are ignored.

For the graph in Figure 4 the subgraphs $\mathcal{G}[\{1\}]$, $\mathcal{G}[\{3\}]$ and $\mathcal{G}[\{2, 4\}]$ are shown in Figures 6(a), (b) and (c) respectively. Note in particular that the edge $2 \rightarrow 3$ is not included in the subgraph $\mathcal{G}[\{2, 4\}]$.

Proposition 2.11. Let \mathcal{G} be an mDAG with districts D_1, \dots, D_k . A probability distribution P with density p is in the marginal model for \mathcal{G} if and only if

$$p(x_V | x_W) = \prod_{i=1}^k g_i(x_{D_i} | x_{\text{pa}(D_i) \setminus D_i}),$$

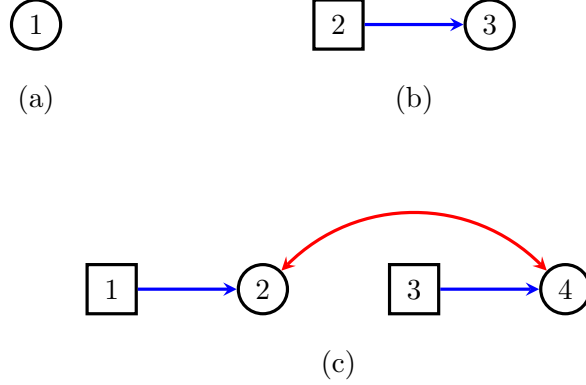


Figure 6: Subgraphs corresponding to factorization of the graph in Figure 4 into districts. Parent nodes of the district are drawn as squares.

where each g_i is a probability density in the marginal model for $\mathcal{G}[D_i]$.

In addition, p is in the marginal model for \mathcal{G} only if for every $v \in \text{sterile}_{\mathcal{G}}(V)$, the marginal distribution

$$p(x_{V \setminus v} | x_W) = \sum_{x_v} p(x_V | x_W)$$

is in the marginal model for \mathcal{G}_{-v} .

Proof. Consider the factorization of the canonical DAG $\bar{\mathcal{G}}$. The first result follows from grouping the factors according to districts and noting that there is no overlap in the variables being integrated out. The second result follows from noting that if v has no children, the variable x_v only appears in a single factor, and this term is a conditional distribution that integrates to 1. \square

This result tells us in particular that we need only consider mDAGs containing a single district, since the characterization of the model can always be reduced to such graphs.

2.3 Relationship between mDAGs and ADMGs

Previous papers considering marginal models for DAGs have used *acyclic directed mixed graphs*, which are the restriction of mDAGs with random vertices so that each bidirected edge has size two (Richardson, 2003; Shpitser et al., 2012; Evans and Richardson, 2014).

From the perspective of the nested Markov property this restriction is not a problem, because if we replace each bidirected hyper-edge in an mDAG with all its subsets of size 2, we reach a conditional ADMG which under the nested Markov property yields the same model. Thus the results of this paper show that, *algebraically*, the model defined by having a single latent

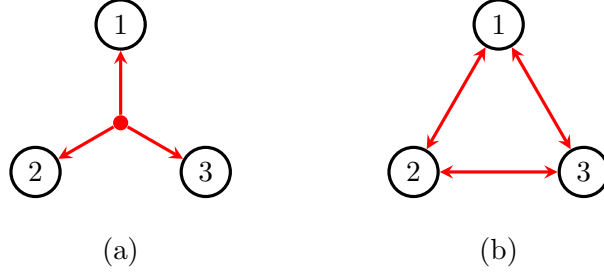


Figure 7: (a) An mDAG on three vertices representing a saturated model; (b) the bidirected 3-cycle, the simplest non-geared mDAG.

parent for several variables is the same as having pairwise parents: contrast the mDAGs in Figure 7, which both represent models of full dimension.

However if we consider the marginal model in full this is false, as the restriction to pairwise latent parents will generally lead to additional inequality constraints. Hence the marginal model for the mDAG in Figure 7(b) is strictly smaller than the one for 7(a) (Fritz, 2012, Proposition 2.13). See Evans (2014) for a more detailed discussion.

3 Nested Markov Property

The *nested Markov property* is defined via constraints satisfied by the marginal model, including conditional independences and ‘dormant independences’ such as the Verma constraint in Example 1.1 (Shpitser et al., 2014). The property is defined in the following recursive way.

Definition 3.1 (Nested Markov Property). A conditional density p obeys the *nested Markov property* for an mDAG $\mathcal{G}(V, W)$ if $V = \emptyset$, or both:

1. p factorizes over the districts D_1, \dots, D_l of \mathcal{G} :

$$p(x_V | x_W) = \prod_{i=1}^l g_i(x_{D_i} | x_{\text{pa}(D_i) \setminus D_i})$$

such that g_i is a distribution obeying the nested Markov property with respect to $\mathcal{G}[D_i]$; and

2. for each $v \in V$ such that $\text{ch}_{\mathcal{G}}(v) = \emptyset$, the marginal distribution

$$p(x_{V \setminus v} | x_W) = \sum_{x_v} p(x_V | x_W)$$

obeys the nested Markov property with respect to $\mathcal{G}(V \setminus \{v\}, W)$, the subgraph induced by removing v .

We denote the set of distributions that obey the nested Markov property with respect to the mDAG \mathcal{G} by $\mathcal{N}(\mathcal{G})$.

Example 3.2. Consider again the mDAG in Figure 4. Applying criterion 1 to this graph implies that

$$p(x_1, x_2, x_3, x_4) = g_1(x_1) \cdot g_{24}(x_2, x_4 | x_1, x_3) \cdot g_3(x_3 | x_2)$$

for some g_1 , g_3 and g_{24} obeying the nested Markov property with respect to the mDAGs in Figures 6(a), (b) and (c) respectively. Applying the second criterion to g_{24} and the now childless vertex 2 (see Figure 6(c)) gives

$$\sum_{x_2} g_{24}(x_2, x_4 | x_1, x_3) = h(x_4 | x_3)$$

for some function h independent of x_1 ; this is precisely the Verma constraint.

The marginal model implies additional conditions on joint distributions, because although it is also closed under marginalization of vertices without children (as in condition 2), this is not sufficient to describe the joint distribution. In particular, for p to be in the marginal model, g_{24} must satisfy Bell's inequalities (see, for example, ver Steeg and Galstyan, 2011).

The nested Markov property is sound with respect to marginal models, in the sense that all constraints represented by the former also hold in the latter. This is formalised in the following result.

Theorem 3.3. *For any mDAG \mathcal{G} we have $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$.*

Proof. This follows from the fact that the nested Markov model is defined in terms of constraints which are proven in Proposition 2.11 to be satisfied by the marginal model. \square

Definition 3.4. Let \mathcal{G} be an mDAG with random vertices V . For an arbitrary set $C \subseteq V$, define $\text{sterile}_{\mathcal{G}}(C) \equiv C \setminus \text{pa}_{\mathcal{G}}(C)$. In words $\text{sterile}_{\mathcal{G}}(C)$ is the subset of C whose elements have no children in C . We say a set C is *sterile* if $C = \text{sterile}_{\mathcal{G}}(C)$.

Let \mathcal{G} be an mDAG. A subset of vertices \mathcal{G} is called *intrinsic* if it is a district in any graph which can be obtained by iteratively applying operations of the form 1 and 2 in Definition 3.1.

Given an intrinsic set, S , define $H = \text{sterile}_{\mathcal{G}}(S)$ to be the *recursive head*, and $T = \text{pa}_{\mathcal{G}}(S)$ the *tail*, associated with S (note that H and T are disjoint). The collection of all recursive heads in \mathcal{G} is denoted $\mathcal{H}(\mathcal{G})$.

Lastly, define

$$\mathcal{A}(\mathcal{G}) \equiv \{H \cup A \mid H \in \mathcal{H}(\mathcal{G}), A \subseteq T\}.$$

to be the *parametrizable sets* of \mathcal{G} .

Example 3.5. The mDAG in Figure 4 has districts $\{1\}$, $\{3\}$ and $\{2, 4\}$, so these are all intrinsic sets. Further, in the subgraph $\mathcal{G}[\{2, 4\}]$ the vertices 2 and 4 have no children, so we can marginalize either to obtain $\{2\}$ and $\{4\}$ as intrinsic sets. The corresponding recursive heads and tails are then:

S	1	2	3	4	2,4
H	1	2	3	4	2,4
T	\emptyset	1	2	3	1,3

Note that intrinsic sets and recursive heads consist only of random vertices, but that tails may include both random and fixed vertices.

Proposition 3.6. *Let C be a bidirected-connected set in an mDAG \mathcal{G} ; then there exists an intrinsic set S such that $C \subseteq S$ and $\text{sterile}_{\mathcal{G}}(S) \subseteq \text{sterile}_{\mathcal{G}}(C)$.*

Proof. The district containing C is intrinsic by definition, so there exists an intrinsic set containing C ; let S be a minimal intrinsic set (by inclusion) containing C . By the definition of intrinsic sets S is a district in some graph reached by iteratively applying the operations 1 and 2 to \mathcal{G} : applying operation 1 again gives the graph $\mathcal{G}[S]$.

Suppose for contradiction that there exists $v \in \text{sterile}_{\mathcal{G}}(S) \setminus \text{sterile}_{\mathcal{G}}(C)$; then $v \notin C$, since otherwise some child of v would be in C , and therefore in S . In addition, v is childless in the subgraph $\mathcal{G}[S]$, so we can remove v under operation 2 of Definition 3.1. In the resulting strictly smaller graph, C is still contained within one district, say S' , since C is bidirected-connected; in addition S' is also intrinsic, so we have found a strictly smaller intrinsic set $S' \supseteq C$, and reached a contradiction. \square

We use the Δ operator to denote the symmetric difference of two sets:

$$A \Delta B \equiv (A \setminus B) \cup (B \setminus A)$$

Given a collection of sets A_i , $i = 1, \dots, k$ indexed by a finite set I , let

$$\bigtriangleup_{i=1}^k A_i \equiv A_1 \Delta A_2 \Delta \dots \Delta A_k.$$

denote the symmetric difference of all the A_i . That is, it is the set containing precisely those elements a which appear in an odd number of the sets A_i .

The following result gives a characterization of the parametrizable sets in terms of symmetric differences which will be fundamental to our proof of the main results in this paper.

Lemma 3.7. *A set $A \in \mathcal{A}(\mathcal{G})$ if and only if there exists a bidirected-connected set $C = \{v_1, \dots, v_k\}$ in \mathcal{G} , and sets A_i , $i = 1, \dots, k$, of the form*

$$\{v_i\} \subseteq A_i \subseteq \{v_i\} \cup \text{pa}_{\mathcal{G}}(v_i),$$

such that

$$A = \bigtriangleup_{i=1}^k A_i = A_1 \triangle \cdots \triangle A_k. \quad (4)$$

Proof. Suppose that $A \in \mathcal{A}(\mathcal{G})$; then $H \subseteq A \subseteq H \cup T$ for some head-tail pair (H, T) , with associated intrinsic set S . Then let $C = S$, since intrinsic sets are by definition bidirected-connected, and consider sets of the form (4). Such a set always contains H , because each $v_i \in H$ appears in A_i and, by sterility, in no other set. Every vertex $t \in T$ is (by definition) the parent of some vertex v_j in S , so we can *choose* whether T appears in a set of the form (4) simply by choosing whether or not to include t in A_j .

Conversely, suppose that A is of the form (4) for some bidirected-connected set C ; let S be an intrinsic set satisfying the conditions of Proposition 3.6, and (H, T) be its associated head-tail pair. Then the head $H = \text{sterile}_{\mathcal{G}}(S) \subseteq \text{sterile}_{\mathcal{G}}(C)$. Each $v_i \in H \subseteq C$ appears in A , since $v_i \in A_j$ if and only if $i = j$. Also $A \subseteq C \cup \text{pa}_{\mathcal{G}}(C) \subseteq S \cup \text{pa}_{\mathcal{G}}(S) = H \cup T$, so $A \in \mathcal{A}(\mathcal{G})$. \square

The following corollary will allow us to generalize our later results to graphs which are not geared.

Corollary 3.8. *Let \mathcal{G} be an mDAG, and $A \in \mathcal{A}(\mathcal{G})$. Then there exists a geared mDAG $\mathcal{G}' \subseteq \mathcal{G}$, such that $A \in \mathcal{A}(\mathcal{G}')$.*

Proof. By Lemma 3.7, A is of the form (4) for some bidirected-connected set C . Let \mathcal{G}' have the same vertices (random and fixed) and directed edges as \mathcal{G} , but be such that the set C is *singly connected* by bidirected edges (i.e. the edges are all of size 2 and removing any of them will cause C to be disconnected) chosen to be a subgraph of \mathcal{G} . Then \mathcal{G}' is geared by standard properties of trees and running intersection, and using Lemma 3.7 again we have $A \in \mathcal{A}(\mathcal{G}')$. \square

3.1 Parametrization of the nested model

The nested Markov model can be parameterized with parameters indexed by head-tail sets, similarly to the ordinary Markov model (Evans and Richardson, 2014); see Evans and Richardson (2015) for details. We now state some consequences of this.

Theorem 3.9. *For a state-space \mathfrak{X}_{VW} the set $\mathcal{N}(\mathcal{G})$ is semi-algebraic, and the variety defined by its Zariski closure is irreducible. Further, the model does not have singularities within the strictly positive probability simplex, and has dimension*

$$d(\mathcal{G}, \mathfrak{X}_{VW}) \equiv \sum_{H \in \mathcal{H}(\mathcal{G})} (|\mathfrak{X}_H| - 1) \cdot |\mathfrak{X}_T|.$$

Proof. That the model is semi-algebraic and has an irreducible Zariski closure follows from the fact that the model can be defined parametrically (see, for example, Cox et al., 2007). The smoothness and dimension are proved in Evans and Richardson (2015). \square

4 Proof Outline

The results in Sections 5 and 6 are fairly technical, so at this point we present a sketch of our approach with a particular example. The main result is proved by showing that tangent cone at the uniform distribution of the marginal model is the same as the tangent space defined by the nested Markov model. To achieve this, we decompose these tangent spaces according to subsets of the vertices in the graph. The following decomposition of the vector space $\mathbb{R}^{|\mathfrak{X}_V|}$ will prove useful.

Definition 4.1. Let Λ_A be the subspace of $\mathbb{R}^{|\mathfrak{X}_V|}$ consisting of vectors p such that

- (i) $\sum_{y_a \in \mathfrak{X}_a} p(y_a, x_{V \setminus \{a\}}) = 0$ for each $a \in A$ and $x_{V \setminus \{a\}} \in \mathfrak{X}_{V \setminus \{a\}}$;
- (ii) $p(x_V) = p(y_V)$ whenever $x_A = y_A$.

In other words, considered as a function $p : \mathfrak{X}_V \rightarrow \mathbb{R}$, the value of p only depends upon x_A , and its sum over x_a for $a \in A$ (keeping the other arguments fixed) is 0. In particular Λ_\emptyset is the subspace spanned by the vector of 1s. In the case where all the variables are binary, each Λ_A corresponds to the space spanned by the corresponding column of a log-linear design matrix.

Proposition 4.2. *The real vector space $\mathbb{R}^{|\mathfrak{X}_V|}$ can be decomposed as the direct sum*

$$\mathbb{R}^{|\mathfrak{X}_V|} = \bigoplus_{A \subseteq V} \Lambda_A.$$

In fact, the spaces Λ_A and Λ_B are orthogonal if $A \neq B$.

Proof. See Appendix A. □

A multivariate model defined by conditional independence constraints always contains the uniform distribution

$$p_0(x_V) \equiv |\mathfrak{X}_V|^{-1}, \quad x_V \in \mathfrak{X}_V,$$

at which point all variables are totally independent. Now, we will show that the tangent cone around of p_0 of all the models we consider is a vector space of the form

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A, \tag{5}$$

for some collection $\mathcal{A}(\mathcal{G})$ of non-empty subsets of V . We will refer, informally, to each of the spaces Λ_A as a ‘direction’, and show that we can perturb the distribution p_0 in any such direction in $\mathcal{A}(\mathcal{G})$. That is, for any vector q in (5), a distribution of the form $p_0 + \eta q + O(\eta^2)$ is contained within the model, so the tangent cone of the model around p_0 contains the vector space Λ_A .

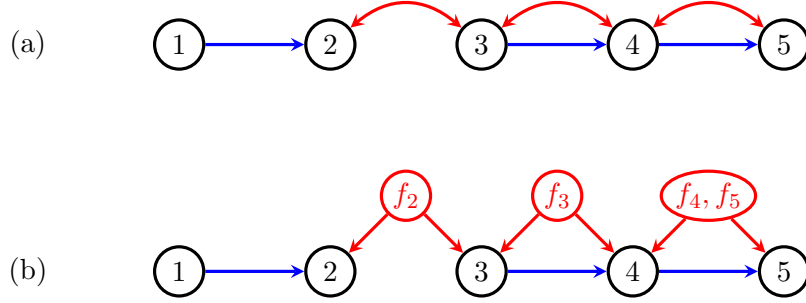


Figure 8: (a) An mDAG, and (b) its canonical DAG with functional latent variables.

4.1 An Example

We will illustrate the main ideas of the proof by considering the graph in Figure 8. One can check that the nested model for this graph is defined by the constraint $X_1 \perp\!\!\!\perp X_3, X_4, X_5$, which means that the model lies within the space orthogonal to $\Lambda_{13} + \Lambda_{14} + \Lambda_{134} + \Lambda_{15} + \Lambda_{135} + \Lambda_{145} + \Lambda_{1345}$. We will show that the associated marginal model can be ‘perturbed’ in every other direction around p_0 .

To achieve this we will fix the state-space of the latent variables to be a series of random functions that (once generated) determine the value of the observed variables. This process is formalized for a large class of graphs in Section 5.

Consider the mDAG in Figure 8(a). The vertex 2 has the observed parent 1 and one latent parent, so we can assume without loss of generality that this latent variable contains a (random) function $f_2 : \mathfrak{X}_1 \rightarrow \mathfrak{X}_2$ which tells X_2 which value it should take, depending upon the value of its parent X_1 . One can show that it is sufficient to fix the state-space of this latent variable to be the finite set of functions $\mathcal{F}_2 \equiv \{f_2 : \mathfrak{X}_1 \rightarrow \mathfrak{X}_2\}$.

Now, we note that the vertex 3 has three parents: 2, the latent vertex now labelled f_2 , and one other latent vertex. Since we have fixed the state-space for all but one latent variable, we can use the same argument to say that without loss of generality the other contains a random function $f_3 : \mathcal{F}_2 \rightarrow \mathfrak{X}_3$ that fixes the value for X_3 given f_2 . In fact we can fix the second latent variable to be the collection of functions $\mathcal{F}_3 = \{f_3 : \mathcal{F}_2 \rightarrow \mathfrak{X}_3\}$.

A similar argument works for X_4 and leads to fixing the state-space of the final latent variable as $\mathcal{F}_4 \times \mathcal{F}_5$, where

$$\mathcal{F}_4 = \{f_4 : \mathfrak{X}_3 \times \mathcal{F}_3 \rightarrow \mathfrak{X}_4\} \quad \mathcal{F}_5 = \{f_5 : \mathfrak{X}_4 \rightarrow \mathfrak{X}_5\};$$

(see Figure 8(b)). We can initially assume that each function is generated by independently and uniformly sampling a value of its output for each combination of values in its input. This will lead to a completely uniform

distribution p_0 over the observed variables. Let TC_0 be the tangent cone of the model $\mathcal{M}(\mathcal{G})$ around p_0 .

Clearly any marginal distribution of X_1 can be obtained, so $\Lambda_1 \subseteq \text{TC}_0$. The function f_2 controls precisely the conditional distribution of X_2 given X_1 , so by manipulating the distribution used to generate f_2 we can obtain any conditional distribution we desire. This shows that the vector space $\Lambda_2 + \Lambda_{12}$ is contained in the tangent cone of our model around p_0 .

The function f_3 controls the distribution of X_3 , so by the same reasoning we can show that $\Lambda_3 \subseteq \text{TC}_0$. However, *in addition* we can change the way X_3 responds to different values of its parent f_2 . Let A_2 be any ‘direction’ that can be perturbed by manipulation of f_2 (i.e. $\{1\}$ or $\{1, 2\}$). We will show (Lemma 6.9) that because f_2 is an argument of the random function f_3 , manipulation of the distribution of f_3 allows us to perturb p_0 in any ‘direction’ A_3 of the form $A_3 = \{3\} \cup A_2$. In other words we can show that $\Lambda_{23} + \Lambda_{123}$ is contained in the tangent cone. Note that the only subset of $\{1, 2, 3\}$ we have not obtained so far is $\{1, 3\}$, and since $X_1 \perp\!\!\!\perp X_3$ under this model we know that this direction *cannot* be perturbed.

By similar reasoning f_4 has f_3 as an argument, so we will be able to manipulate the sets $\{3, 4\}$, $\{2, 3, 4\}$ and $\{1, 2, 3, 4\}$. However f_4 also has the argument X_3 , meaning that it can control the conditional distribution of X_4 given X_3 , and hence the directions $\{4\}$ and $\{3, 4\}$. By combining the dependence upon these two variables we will show that in fact we can push in any direction of the form $A_4 = \{4\} \triangle A_3$ or $A_4 = \{3, 4\} \triangle A_3$, which, all told will allow us to obtain any of the directions in $\Lambda_4 + \Lambda_{24} + \Lambda_{34} + \Lambda_{124} + \Lambda_{234} + \Lambda_{1234}$.

Finally, the function f_5 controls the conditional distribution of X_5 given X_4 , so by manipulating its distribution we can obtain the $\Lambda_5 + \Lambda_{45}$ directions. However because we can alter the joint distribution of (f_4, f_5) in an arbitrary way, we can actually obtain any direction of the form $A_4 \triangle A_5$, giving the additional sets

$$25 \quad 125 \quad 35 \quad 235 \quad 1235 \quad 245 \quad 1245 \quad 345 \quad 2345 \quad 12345.$$

All directions are now accounted for, and since they are all obtained by local perturbations of a particular parameterization, in fact the different directions form a vector space; the tangent cone TC_0 is therefore the tangent space orthogonal to $\Lambda_{13} + \Lambda_{14} + \Lambda_{134} + \Lambda_{15} + \Lambda_{135} + \Lambda_{145} + \Lambda_{1345}$ and the marginal model is locally equivalent to the model of independence $X_1 \perp\!\!\!\perp X_3, X_4, X_5$.

5 Geared mDAGs

We define a special class of mDAGs which we term ‘geared’. For such graphs, the state-space of the hidden vertices can be restricted without loss of generality, making proofs concerning the marginal model considerably easier.

Definition 5.1. Let \mathcal{G} be an mDAG with bidirected hyper-edge set \mathcal{B} . We say that \mathcal{G} is *geared* if the elements of \mathcal{B} satisfy the running intersection property. That is, there is an ordering of the edges B_1, \dots, B_k such that for each $j > 1$, there exists $s(j) < j$ with

$$B_j \cap \bigcup_{i < j} B_i = B_j \cap B_{s(j)}.$$

In other words, all the vertices B_j shares with any previous edge are contained within one such edge.

A particular ordering of the elements of \mathcal{B} which satisfies running intersection is called a *gearing* of \mathcal{G} .

The term ‘geared’ is chosen because a collection of bidirected edges which satisfies running intersection may appear rather like ‘cogs’ in a set of gears: see Figure 5. The definition is very similar to the idea of decomposability in an undirected graph; however we avoid using this terminology, because DAGs (which have no bidirected edges are therefore trivially geared) may or may not be decomposable in the original sense (Lauritzen, 1996).

Example 5.2. The simplest non-geared mDAG is the bidirected 3-cycle, depicted in Figure 7(b); there is no way to order the bidirected edge sets $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$ in a way which satisfies the running intersection property, since whichever edge is placed last in the ordering shares a different vertex with each of the two other edges.

Given a single-district, geared mDAG with at least one bidirected edge and a gearing B_1, \dots, B_k , define

$$R_j = B_j \setminus \bigcup_{i < j} B_i$$

(taking $R_1 = B_1$). We call R_j the *remainder set* associated with B_j , and the remainder sets partition the random vertices V . In addition, for a random vertex $v \in V$, define $r(v)$ to be the unique j such that $v \in R_j$.

Now say that an ordering $<$ on the vertices in V *respects the gearing* if for $v \in R_i$ and $w \in R_j$, we have $v < w$ whenever $i > j$; in other words, all the vertices in R_k precede all those in R_{k-1} , etc; such an ordering always exists.

For each $v \in V$ with $r(v) = j$, define

$$\pi(v) = \bigcup_{\substack{i > j \\ v \in B_i}} R_i;$$

that is, the remainders associated with all bidirected edges which contain v and are later than j in the ordering. Then define a collection of functions

$$\mathcal{F}_v \equiv \{f : \mathfrak{X}_{\text{pa}(v)} \times \mathcal{F}_{\pi(v)} \rightarrow \mathfrak{X}_v\},$$

where $\mathcal{F}_A = \times_{a \in A} \mathcal{F}_a$ and $\mathcal{F}_\emptyset = \mathfrak{X}_\emptyset = \{1\}$. This is valid recursive definition, since all the vertices in $\pi(v)$ precede v in an ordering which respects the gearing.

Example 5.3. Consider the mDAG in Figure 5, and order the bidirected edges as

$$B_1 = \{1, 2\}, \quad B_2 = \{2, 3, 4\}, \quad B_3 = \{3, 4, 5\}$$

giving remainder sets

$$R_1 = \{1, 2\}, \quad R_2 = \{3, 4\}, \quad R_3 = \{5\}.$$

The ordering $5 < 4 < 3 < 2 < 1$ of the random vertices respects the gearing, and we have

$$\pi(1) = \pi(5) = \emptyset, \quad \pi(3) = \pi(4) = \{5\}, \quad \pi(2) = \{3, 4\}.$$

In this case then

$$\begin{aligned} \mathcal{F}_5 &= \{f : \mathfrak{X}_3 \rightarrow \mathfrak{X}_5\} \\ \mathcal{F}_4 &= \{f : \mathfrak{X}_{2,3,6} \times \mathcal{F}_5 \rightarrow \mathfrak{X}_4\} \\ \mathcal{F}_3 &= \{f : \mathfrak{X}_1 \times \mathcal{F}_5 \rightarrow \mathfrak{X}_3\} \\ \mathcal{F}_2 &= \{f : \mathcal{F}_{3,4} \rightarrow \mathfrak{X}_2\} \\ \mathcal{F}_1 &= \{f : \{1\} \rightarrow \mathfrak{X}_1\} \quad (\text{or equivalently } \mathcal{F}_1 = \mathfrak{X}_1). \end{aligned}$$

Alternatively, if we order the bidirected edges as $\{2, 3, 4\}$, $\{1, 2\}$, $\{3, 4, 5\}$, then we could take $1 < 5 < 2 < 3 < 4$, and

$$\pi(1) = \pi(5) = \emptyset, \quad \pi(3) = \pi(4) = \{5\}, \quad \pi(2) = \{1\};$$

this yields $\mathcal{F}_2 = \{f : \mathcal{F}_1 \rightarrow \mathfrak{X}_2\}$, with other collections \mathcal{F}_v remaining unchanged.

5.1 Functional Models

The property that makes geared graphs useful is that we can find a latent variable model with all variables discrete that yields the same set of distributions over the observed variables as the marginal model. This fact provides a tool with which to attack the main result of this paper, and demonstrate the true dimension of mDAG models.

If a vertex v is contained within exactly one bidirected edge, B , then without loss of generality we can assume that the latent variable corresponding to B contains all the residual information about how X_v should behave given the values of its visible parents, $X_{\text{pa}(v)}$. In other words, the latent variable associated with B contains a (random) function $f_v : \mathfrak{X}_{\text{pa}(v)} \rightarrow \mathfrak{X}_v$ which ‘tells’ $X_v = f_v(X_{\text{pa}(v)})$ which value it should take for each value of its other parents.

However, if v is contained within two or more bidirected edges, say B_i and B_j , it is not clear how to define such a function until the state-space associated with one of these latent parents has already been fixed. The decomposable structure of geared graphs makes it possible to iteratively fix finite state-spaces for each bidirected edge without loss of generality.

Specifically, for a single-district, geared mDAG \mathcal{G} with remainder sets R_1, \dots, R_k , first form the canonical DAG $\bar{\mathcal{G}}$ by replacing each bidirected edge B_i in \mathcal{G} with a new vertex u_i , such that $\text{ch}_{\bar{\mathcal{G}}}(u_i) = B_i$. Compare, for example, the structure of the graphs in Figures 5(a) and (b). Then define independent latent variables

$$U_i \equiv (f_v \in \mathcal{F}_v \mid v \in R_i), \quad i = 1, \dots, k,$$

For example, with the first gearing given in Example 5.3 for the graph in Figure 5(a), we would have

$$U_1 = (f_1, f_2), \quad U_2 = (f_3, f_4), \quad U_3 = (f_5).$$

Associating each variable U_i with the vertex u_i leads to the DAG in Figure 9. Notice that, for each $v \in V$, the function f_v is contained within a parent variable of v . In addition, all the arguments of the function f_v are also parents of v .

For example, take $v = 4$, and note that $f_4 \in \mathcal{F}_4$ is determined from $U_2 = (f_3, f_4)$, and the associated vertex u_2 a parent of 4. In addition, $\mathcal{F}_4 = \{f : \mathfrak{X}_{2,3,6} \times \mathcal{F}_5 \rightarrow \mathfrak{X}_4\}$, so the arguments of the function f_4 , namely X_2, X_3, X_6 and f_5 , all correspond to vertices which are also parents of 4 in Figure 9 (see Figure 10). Thus, in setting $X_4 = f_4(X_2, X_3, X_6, f_5)$ we ensure that X_4 is a well defined function of its parent variables.

In fact using this construction we can set

$$X_v = f_v(f_{\pi(v)}, X_{\text{pa}(v)})$$

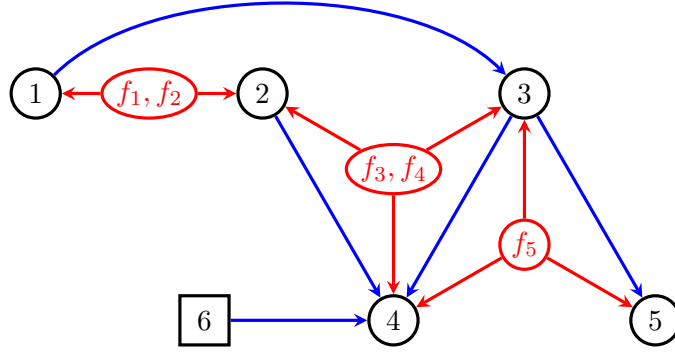


Figure 9: A DAG with functional latent variables, associated with a gearing of the mDAG in Figure 5(a).

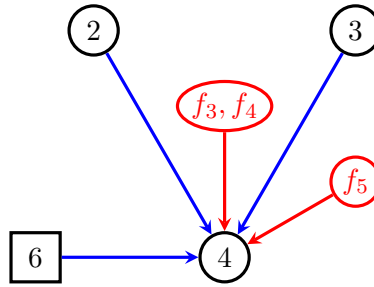


Figure 10: Subgraph of the DAG in Figure 9 containing the vertex 4 and its parents.

for every $v \in V$, which is well defined because the directed part of the original mDAG is acyclic. The following result shows that the resulting conditional distribution over X_V given X_W is in the marginal model for the original mDAG.

Theorem 5.4. *Let \mathcal{G} be a geared mDAG, and $R_i, i = 1, \dots, k$ be the remainder sets corresponding to some gearing of \mathcal{G} . Suppose we generate functions $f_v \in \mathcal{F}_v$ according to a distribution in which*

$$(f_v \mid v \in R_i) \perp\!\!\!\perp (f_w \mid w \in V \setminus R_i),$$

for each $i = 1, \dots, k$, and then define

$$X_v = f_v(f_{\pi(v)}, X_{\text{pa}(v)}), \quad v \in V.$$

Then the induced conditional distribution P on X_V given X_W is in the marginal model for \mathcal{G} .

Proof. For each bidirected edge B_i , define the random variable $U_i = (f_v \mid v \in R_i)$. The U_i s are represented by exogenous variables on the DAG $\bar{\mathcal{G}}$, and the conditions given in the statement of the theorem ensures they are all independent. The structural equation property for $\bar{\mathcal{G}}$ will therefore be satisfied if each X_v is a well defined function of its parents in the graph.

In other words, the three components f_v , $f_{\pi(v)}$ and $X_{\text{pa}(v)}$ must all be determined from random variables which are parents of v in $\bar{\mathcal{G}}$. This holds for $X_{\text{pa}(v)}$ by definition. Additionally $v \in R_i$ implies that $v \in B_i$, and that therefore the variable U_i is a parent variable of X_v ; then since the function f_v is just a component of U_i , this is indeed determined by a parent of v .

Lastly suppose $w \in \pi(v)$; this happens if and only if $w, v \in B_j$ for some $j > i$, in which case $w \in R_j$ for the minimal such j by the running intersection property of the gearing. Then f_w is contained in U_j , which is also a parent variable of X_v .

Thus f_v , $f_{\pi(v)}$ and $X_{\text{pa}(v)}$ are all well defined functions of parent variables of v , and so setting $X_v = f_v(f_{\pi(v)}, X_{\text{pa}(v)})$ respects the Markov property of the graph. \square

The idea of this formulation is that f_v is a random function that ‘tells X_v what to do,’ or rather what value to take, given the values of its other parents. If some of those other parents are also latent, then they must be defined first, and the need to do this in a well-ordered manner explains why it is necessary for \mathcal{G} to be geared.

In fact it follows from a slight variation of Proposition 5.2 in Evans (2014) that *any* distribution in the marginal model of a geared graph can be generated in the way described in Theorem 5.4. Since each of these latent variables takes values in a finite collection of functions, this means that the marginal model is equivalent to the margin of a Bayesian network in which

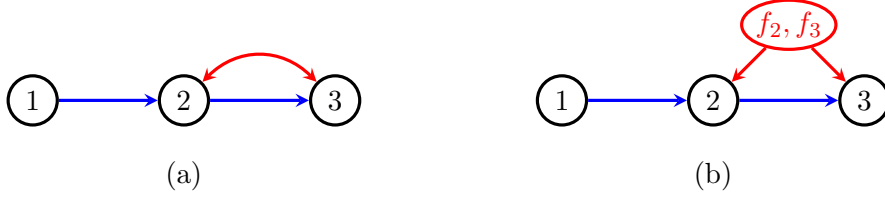


Figure 11: (a) An mDAG representing the instrumental variables model; (b) a DAG with functional latent variables equivalent to the potential outcomes model of instrumental variables.

all the random variables (latent and observed) are finite and discrete. It follows that marginal models for geared mDAGs are semi-algebraic sets.

Example 5.5. Consider the mDAG in Figure 11(a), which represents the *instrumental variables* model. This is used, for example, to model non-compliance in clinical trials, with X_1 representing a randomized treatment, X_2 the treatment actually taken, and X_3 a clinical outcome.

This mDAG has only one bidirected edge and therefore is trivially geared with $R_1 = B_1 = \{2, 3\}$. This leads to functional latent variables f_2 and f_3 , where

$$\begin{aligned}\mathcal{F}_2 &= \{f : \mathfrak{X}_1 \rightarrow \mathfrak{X}_2\} \\ \mathcal{F}_3 &= \{f : \mathfrak{X}_2 \rightarrow \mathfrak{X}_3\}.\end{aligned}$$

The resulting DAG model is shown in Figure 11(b). The function f_2 defines, for an individual, which treatment she actually takes given which arm of the trial she is assigned to; this is known as her *compliance type*. Similarly f_3 determines what the patient's outcome will be given each possible treatment she chooses to take, known as her *response type*. These functions are precisely the *potential outcomes* of the Neyman-Rubin causal framework (Neyman, 1923; Rubin, 1974; Richardson et al., 2011).

We note that for our purposes the functions f_v are a purely mathematical construct, and thus philosophical questions about the nature and existence of potential outcomes have no direct bearing on the results herein (see, for example, Dawid, 2000).

6 Main Results

In this section we provide the technical results to prove the main theorem. This is done first for geared mDAGs, and the result is then extended to general graphs.

6.1 Distributions for Geared mDAGs

Let \mathcal{G} be a single-district, geared mDAG, with gearing given by remainder sets R_1, \dots, R_k ; assign a probability distribution ρ_i to each collection of functions $U_i \equiv (f_v | v \in R_i)$. Suppose we draw variables $U_i = (f_v)_{v \in R_i}$ independently according to ρ_i , and use them to generate observed variables X_V for each possible value of the fixed vertices X_W . Applying Theorem 5.4, the resulting (conditional) distribution over the observed variables, say P , is in the marginal model for \mathcal{G} .

Let $\pi(R_i) \equiv \bigcup_{v \in R_i} \pi(v)$ and $f_A \equiv (f_v | v \in A)$. Define

$$p[\rho_k, \dots, \rho_1](x_V | x_W) = \sum_{\Phi_k(x_{VW})} \rho_k(f_{R_k}) \cdots \sum_{\Phi_1(f_{\pi(R_1)}, x_{VW})} \rho_1(f_{R_1}),$$

Where

$$\Phi_i(f_{\pi(R_i)}, x_{VW}) = \{f_{R_i} | f_v(x_{\text{pa}(v)}, f_{\pi(v)}) = x_v \text{ for each } v \in R_i\}; \quad (6)$$

that is, Φ_i gives us precisely the set of functions f_{R_i} that, given appropriate values of parents variables, jointly evaluate to x_{R_i} . Ultimately, then, we have a sum over combinations of functions f_V that, given the input $X_W = x_W$, jointly evaluate to x_V .

The function $p[\cdot]$ returns a vector indexed by x_{VW} representing the induced conditional probability distribution of X_V given X_W . For brevity we will generally denote this as

$$p[\rho_k, \dots, \rho_1] = \sum_{\Phi_k} \rho_k \cdots \sum_{\Phi_1} \rho_1,$$

with the dependence upon x_{VW} left implicit. It may be helpful to think of this as a family of probability distributions for X_V given X_W , indexed by parameters ρ_1, \dots, ρ_k .

Example 6.1. In the case of the mDAG in Figure 5(a) we have three bidirected edges and remainder sets, and the gearing used in Figure 9 gives

$$p[\rho_3, \rho_2, \rho_1] = \sum_{\Phi_3} \rho_3(f_5) \sum_{\Phi_2} \rho_2(f_3, f_4) \sum_{\Phi_1} \rho_1(f_1, f_2),$$

where

$$\begin{aligned} \Phi_1 &= \{(f_1, f_2) : f_1 = x_1, f_2(f_3, f_4) = x_2\} \\ \Phi_2 &= \{(f_3, f_4) : f_3(x_1) = x_3, f_4(x_2, x_3, x_6, f_5) = x_4\} \\ \Phi_3 &= \{f_5 : f_5(x_3) = x_5\}. \end{aligned}$$

By Theorem 5.4, for any ρ_1, \dots, ρ_k , the induced distribution $p[\rho_k, \dots, \rho_1]$ on \mathfrak{X}_V given \mathfrak{X}_W is in the marginal model for \mathcal{G} .

It is clear that choosing $\rho_i(f_{R_i}) = 1$ for each i (up to a constant of proportionality which, for simplicity, we ignore) induces the uniform distribution, p_0 , on \mathfrak{X}_V for each $x_W \in \mathfrak{X}_W$. In other words, the uniform distribution $p_0 \equiv p[1, \dots, 1]$ is contained within $\mathcal{M}(\mathcal{G})$ for any mDAG \mathcal{G} .

6.2 Tangent cones

Definition 6.2. Let \mathfrak{A} be a subset of \mathbb{R}^k containing a point \mathbf{x} . The *tangent cone* of \mathfrak{A} at \mathbf{x} is the set of vectors \mathbf{v} which are of the form

$$\mathbf{v} = \lim_{n \rightarrow \infty} \eta_n^{-1}(\mathbf{v}_n - \mathbf{x})$$

where $\eta_n \rightarrow 0$ and each $\mathbf{v}_n \in \mathfrak{A}$.

A tangent cone is a cone, but may or may not be a vector space, depending upon whether the set \mathfrak{A} is regular at \mathbf{x} . We claim here, though will not need to prove directly, that the tangent cone of $\mathcal{N}(\mathcal{G})$ around the uniform distribution p_0 is the vector space

$$\text{TS}_0^n \equiv \bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A.$$

In fact we will show that this vector space is equal to TC_0 , the tangent cone of $\mathcal{M}(\mathcal{G})$ at the uniform distribution, and the characterization of TS_0^n given here will follow from the fact that $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$ and dimension counting.

Definition 6.3. Let $\lambda : \mathfrak{X}_A \rightarrow \mathbb{R}$; we say that λ is *A-degenerate* (or just degenerate) if for each $a \in A$, and $x_{A \setminus a} \in \mathfrak{X}_{A \setminus a}$,

$$\sum_{y_a} \lambda(y_a, x_{A \setminus a}) = 0.$$

It is clear that the set of *A*-degenerate functions is isomorphic to the vector space Λ_A , though both formulations will be useful.

The main result of this section follows.

Theorem 6.4. *The tangent cone of $\mathcal{M}(\mathcal{G})$ around p_0 is the vector space*

$$\text{TC}_0 = \bigoplus_{A \in \mathcal{A}} \Lambda_A.$$

The proof is delayed until the end of the section. We note that the tangent cone is a vector space, and it has the same dimension as the nested model.

6.3 Results for Geared Graphs

Definition 6.5. Given a degenerate function $\varepsilon_i : \mathcal{F}_{R_i} \rightarrow \mathbb{R}$, define

$$D_i(\varepsilon_i) = \lim_{\eta \downarrow 0} \eta^{-1} \{p[1, \dots, 1 + \eta\varepsilon_i, \dots, 1] - p[1, \dots, 1, \dots, 1]\},$$

so that $D_i(\varepsilon_i)$ is a vector in $\mathbb{R}^{|\mathfrak{X}_V|}$. For sufficiently small $\eta > 0$, $1 + \eta\varepsilon_i$ is non-negative and therefore a valid distribution over \mathcal{F}_{R_i} ; it follows that $D_i(\varepsilon_i) \in \text{TC}_0(\mathcal{G})$, the tangent cone of $\mathcal{M}(\mathcal{G})$ at p_0 .

Let

$$T_i = \{D_i(\varepsilon_i) \mid \varepsilon_i \text{ degenerate}\}.$$

Then T_i is a vector space, since the function $p[\cdot]$ is differentiable at (ρ_k, \dots, ρ_1) , and $T_1 + \dots + T_k$ is contained within the tangent cone of \mathcal{M} around the uniform distribution.

It will be useful to define the following collection of supersets of Φ_i , for $B \subseteq V$:

$$\Phi_i^B(f_{\pi(R_i)}, x_{VW}) \equiv \{f_{R_i} \mid f_v(x_{\text{pa}(v)}, f_{\pi(v)}) = x_v \text{ for each } v \in R_i \cap B\}. \quad (7)$$

Lemma 6.6. Let $C \subseteq R_i$, with $\text{sterile}_G(C) \subseteq A \subseteq C \cup \text{pa}_G(C)$ and $E \subseteq \pi(C)$. Then for every degenerate function

$$\lambda : \mathfrak{X}_A \times \mathcal{F}_E \rightarrow \mathbb{R},$$

there exists a degenerate function $\delta : \mathcal{F}_C \rightarrow \mathbb{R}$ such that

$$\sum_{f_{R_i} \in \Phi_i} \delta(f_C) = \lambda(x_A, f_E),$$

where Φ_i is given by (6). In addition,

$$\sum_{f_{R_i} \in \Phi_i^B} \delta(f_C) = \begin{cases} |\mathfrak{X}_{R_i \setminus B}| \lambda(x_A, f_E) & \text{if } C \subseteq B \\ 0 & \text{otherwise.} \end{cases}$$

Proof. See appendix, Section A.3. □

Remark 6.7. Note that if we set $E = \emptyset$, the above result shows that for appropriate λ and δ ,

$$\begin{aligned} & \eta^{-1} \{p[1, \dots, 1 + \eta\delta, \dots, 1] - p[1, \dots, 1, \dots, 1]\} \\ &= \eta^{-1} \left\{ \sum_{\Phi_k} \dots \sum_{\Phi_i} \eta\delta(f_C) \sum_{\Phi_{i-1}} \dots \sum_{\Phi_1} 1 \right\} \\ &\propto \lambda. \end{aligned}$$

Hence $\Lambda_A \leq T_i$ (i.e. Λ_A is a subspace of T_i) for any A such that $\text{sterile}_{\mathcal{G}}(C) \subseteq A \subseteq C \cup \text{pa}_{\mathcal{G}}(C)$ and $C \subseteq R_i$.

This tells us that we can obtain certain directions in our model's tangent space by only manipulating the distribution of a single latent variable, U_i . For the full range of this space to be achieved it will be necessary to manipulate the distribution of several adjacent⁴ latent variables in a co-ordinated way.

We will extend the previous result to bidirected-connected sets which span multiple remainder sets, though we need the following lemma to ensure that distribution over our sum works as expected.

Lemma 6.8. *Let \mathcal{G} be a single-district, geared mDAG, and C a bidirected-connected set of vertices. We can construct a rooted tree Π_C with vertex set*

$$I_C = \{i \mid R_i \cap C \neq \emptyset\},$$

and such that $i \rightarrow j$ only if there exist $v_j \in R_j \cap C$ and $v_i \in R_i \cap C$ such that $v_j \in \pi(v_i)$.

Proof. See appendix, Section A.4. □

The next result forms the backbone for proving Theorem 6.4: it extends Lemma 6.6 to sets C which may not be contained within a single remainder set.

Lemma 6.9. *Let C be a bidirected-connected set, and define $C_i \equiv C \cap R_i$ and $I \equiv \{i \mid C_i \neq \emptyset\}$. For $\text{sterile}_{\mathcal{G}}(C_i) \subseteq A_i \subseteq C_i \cup \text{pa}_{\mathcal{G}}(C_i)$, let*

$$A = \bigtriangleup_{i \in I} A_i.$$

Then $\Lambda_A \leq T_l$, where l is the minimal element of I .

Proof. By Lemma 6.8, there exists a rooted tree Π with vertices I , such that $i \rightarrow j$ in Π only if there exist $v_i \in R_i \cap C$ and $v_j \in R_j \cap C$ with $v_j \in \pi(v_i)$. In particular $i \rightarrow j$ only if $C_j \subseteq \pi(v_i)$.

Let l be the root node of Π , and for each $j \in \text{ch}_{\Pi}(l)$ denote by Π_j the rooted tree with root j formed only from the descendants of j .

Let $\lambda_i : \mathfrak{X}_{A_i} \rightarrow \mathbb{R}$ be arbitrary A_i -degenerate functions for each $i \in I$. Then starting with vertices which have no children (i.e. the leaves of the tree), and using Lemma 6.6, recursively define δ_i for $i \in I$ as the degenerate function of f_{C_i} such that

$$\sum_{\Phi_i} \delta_i(f_{C_i}) = \lambda_i(x_{A_i}) \prod_{j \in \text{ch}_{\Pi}(i)} \delta_j(f_{C_j}),$$

⁴Adjacent in the sense that they share an observable child vertex.

where the empty product is defined to be equal to 1. Then

$$\sum_{\Phi_k} \cdots \sum_{\Phi_{l+1}} \sum_{\Phi_l} \delta_l(f_{C_l}) \propto \lambda_l(x_{A_l}) \sum_{\Phi_k} \cdots \sum_{\Phi_{l+1}} \prod_{j \in \text{ch}_\Pi(l)} \delta_j(f_{C_j}).$$

For each $i \in I$ an expression of the form $\sum_{\Phi_i} \delta_i(f_{C_i})$ is only a function of f_{C_α} for $\alpha \in \text{ch}_\Pi(i)$ and Π is a tree, so the sum factorizes into components only involving the descendants of each $j \in \text{ch}_\Pi(l)$:

$$\sum_{\Phi_k} \cdots \sum_{\Phi_1} \delta_l(f_{C_l}) \propto \lambda_l(x_{A_l}) \prod_{j \in \text{ch}_\Pi(l)} \sum_{\substack{\Phi_s \\ s \in \text{de}_\Pi(j)}} \delta_j(f_{C_j}).$$

But then for each j the factor represents a disjoint sub-tree Π_j with root node j , so we can just iterate this process within each factor, and get

$$\propto \prod_{i \in I} \lambda_i(x_{A_i}).$$

It follows that any function of the form

$$\prod_{i \in I} \lambda_i(x_{A_i})$$

lies in T_l ; since Λ_A is spanned by such functions it then follows from Lemma A.3 (see Appendix A) that $\Lambda_A \leq T_l$. \square

Corollary 6.10. *For geared graphs \mathcal{G} , we have*

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq T_1 + \cdots + T_k.$$

Proof. Reformulating Lemma 3.7 slightly, for any $A \in \mathcal{A}(\mathcal{G})$ there exists a bidirected-connected set $C = \bigcup_i C_i = \bigcup_i \{v_{i1}, \dots, v_{ik_i}\}$, where $C_i = C \cap R_i$ (we have changed nothing other than to label the vertices v_{ij} by which remainder set they are contained in). Then A is of the form

$$A = \bigtriangleup_{i,j} A_i^j = \bigtriangleup_i \left(\bigtriangleup_j A_i^j \right)$$

for some sets A_i^j such that $\{v_{ij}\} \subseteq A_i^j \subseteq \{v_{ij}\} \cup \text{pa}_{\mathcal{G}}(v_{ij})$.

Applying Lemma 3.7 in reverse to the bidirected-connected set C_i shows that $A_i \equiv \bigtriangleup_j A_i^j$ is in $\mathcal{A}(\mathcal{G})$, and therefore satisfies $\text{sterile}_{\mathcal{G}}(C_i) \subseteq A_i \subseteq C_i \cup \text{pa}_{\mathcal{G}}(C_i)$. Then by Lemma 6.9 the space Λ_A is contained in some T_i , $i = 1, \dots, k$. \square

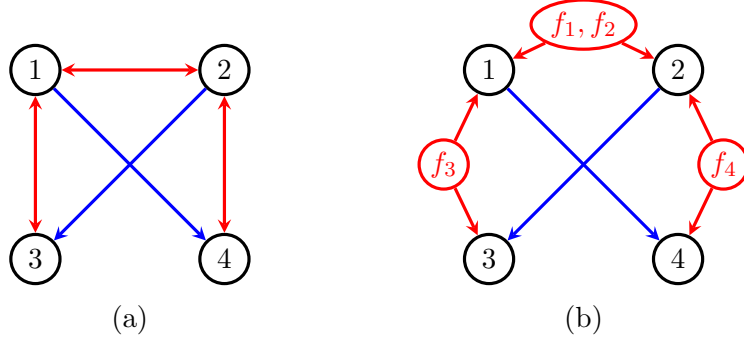


Figure 12: (a) an mDAG on 4 variables, and (b) a DAG with hidden variables corresponding to a gearing of the mDAG in (a).

Example 6.11. Consider the single-district, geared mDAG in Figure 12(a); the nested Markov model for this graph is saturated, and thus $\mathcal{A}(\mathcal{G})$ consists of all non-empty subsets of $\{1, 2, 3, 4\}$.

H	1	2	12	3	13	4	24	34
T	\emptyset	\emptyset	\emptyset	2	2	1	1	12
\mathcal{A}	1	2	12	3, 23	13, 123	4, 14	24, 124	34, 134, 234, 1234

Consider the gearing

i	B_i	R_i
1	$\{1, 2\}$	$\{1, 2\}$
2	$\{1, 3\}$	$\{3\}$
3	$\{2, 4\}$	$\{4\}$

and ordering $3 < 4 < 1 < 2$ which respects this gearing. This leads to the hidden variable model in Figure 12(b); here

$$\begin{aligned}
 f_3 : \mathfrak{X}_2 &\rightarrow \mathfrak{X}_3 & f_4 : \mathfrak{X}_1 &\rightarrow \mathfrak{X}_4 \\
 f_1 : \mathcal{F}_3 &\rightarrow \mathfrak{X}_1 & f_2 : \mathcal{F}_4 &\rightarrow \mathfrak{X}_2.
 \end{aligned}$$

Applying Lemma 6.6 to each remainder set in turn tells us that

$$\begin{aligned}
 \Lambda_1 + \Lambda_2 + \Lambda_{12} &\leq T_1 \\
 \Lambda_3 + \Lambda_{23} &\leq T_2 \\
 \Lambda_4 + \Lambda_{14} &\leq T_3.
 \end{aligned}$$

We can apply Lemma 6.9 with the connected set $C = \{1, 2, 3, 4\}$ to find that $\Lambda_A \leq T_1$, where A is of the form $A = \{1, 2\} \triangle A_2 \triangle A_3$ and

$$\{3\} \subseteq A_2 \subseteq \{2, 3\} \quad \{4\} \subseteq A_3 \subseteq \{1, 4\};$$

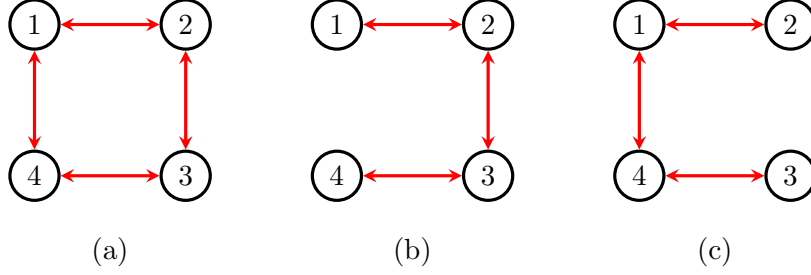


Figure 13: (a) the bidirected 4-cycle, and (b), (c) two geared subgraphs.

that is, $A \in \{\{3, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$, and so

$$\Lambda_{34} + \Lambda_{134} + \Lambda_{234} + \Lambda_{1234} \leq T_1.$$

Repeating with $C = \{1, 2, 3\}$ and $\{1, 2, 4\}$ respectively gives $\Lambda_{13} + \Lambda_{123} \leq T_1$ and $\Lambda_{24} + \Lambda_{124} \leq T_1$.

Thus for every non-empty $A \subseteq \{1, 2, 3, 4\}$ there is some $i \in \{1, 2, 3\}$ such that $\Lambda_A \leq T_i$, and therefore the tangent cone of $\mathcal{M}(\mathcal{G})$ around the uniform distribution is the same as that of the saturated model on four variables. In other words the nested model and marginal model are both of full dimension.

Evans (2012) shows that the marginal model associated with this graph induces some inequality constraints on the joint distribution, and so the nested and marginal models are not identical.

6.4 Dealing with non-geared graphs

Corollary 6.10 put us in a position to prove Theorem 6.4 for geared graphs; however it does not so far extend to the general case, because we cannot fix the state-spaces of the latent variables without a gearing. In this section we will show that the tangent cone of a general marginal model around the uniform distribution is just composed of the tangent cones of its geared subgraphs, and that therefore the problem can be reduced to geared graphs.

Proposition 6.12. *Let \mathcal{G} be an arbitrary mDAG containing geared subgraphs $\mathcal{G}_1, \dots, \mathcal{G}_k$. Suppose that, for each subgraph and a suitable gearing $\Lambda_{A_i} \leq \text{TC}_0(\mathcal{G}_i)$ as a consequence of the earlier results in this section. Then $\Lambda_{A_1} + \dots + \Lambda_{A_k} \leq \text{TC}_0(\mathcal{G})$*

In other words, the tangent cone of \mathcal{G} includes the vector space spanned by all the tangent cones of the subgraphs.

Proof. First consider the case $W = \emptyset$ and $k = 2$, from which the general result will follow similarly.

Let $p_1 \in \mathcal{M}(\mathcal{G}_1) \subseteq \mathcal{M}(\mathcal{G})$ be formed by random functions f_V according to a gearing of \mathcal{G}_1 , and $p_2 \in \mathcal{M}(\mathcal{G}_2) \subseteq \mathcal{M}(\mathcal{G})$ by random functions \tilde{f}_V . Let U_v be independent Bernoulli($\frac{1}{2}$) variables, and define a new distribution by setting

$$Z_v = U_v f_v(f_{\pi(v)}, Z_{\text{pa}(v)}) + (1 - U_v) \tilde{f}_v(\tilde{f}_{\pi(v)}, Z_{\text{pa}(v)});$$

i.e. we randomly (and independently of all other vertices) choose one of the mechanisms f_v or \tilde{f}_v to generate Z_v . Note that although f_v and \tilde{f}_v are independent the values of $f_v(f_{\pi(v)}, Z_{\text{pa}(v)})$ and $\tilde{f}_v(\tilde{f}_{\pi(v)}, Z_{\text{pa}(v)})$ are not, since they share parent variables.

Denote the resulting joint distribution of Z_V by p . It is clear that $p \in \mathcal{M}(\mathcal{G})$, since we are still generating each variable as a random function of its parents and some independent noise, which clearly satisfies the Markov property for $\bar{\mathcal{G}}$.

Now place a distribution over f_V which is uniform except for a perturbation $\eta \delta(f_{C_i})$ which leads to a perturbation $\eta \lambda_{A_1}(x_{A_1})$ over the observed joint distribution for X_V . Similarly for \tilde{f}_V . Then

Then we have

$$\begin{aligned} P(Z_V = z_V) &= \sum_{B \subseteq V} P(U_B = 1, U_{V \setminus B} = 0, X_B = z_B, Y_{V \setminus B} = z_{V \setminus B}) \\ &= \frac{1}{2^{|V|}} \sum_{B \subseteq V} P(X_B = z_B, Y_{V \setminus B} = z_{V \setminus B}). \end{aligned}$$

It follows from the proof of Lemma 6.9 that if $A \in \mathcal{A}(\mathcal{G})$ and $\lambda_A \in \Lambda_A$ then there exists a degenerate $\delta(f_{C_i})$ such that

$$\sum_{\Phi_k} \cdots \sum_{\Phi_i} \delta(f_{C_i}) \cdots \sum_{\Phi_1} 1 = \lambda_A(x_A)$$

and from Lemma 6.6 that

$$\sum_{\Phi_k^B} \cdots \sum_{\Phi_i^B} \delta(f_{C_i}) \cdots \sum_{\Phi_1^B} 1 = \begin{cases} |\mathfrak{X}_{V \setminus B}| \lambda_A(x_A) & \text{if } C \subseteq B \\ 0 & \text{otherwise.} \end{cases}$$

Now since the functions used to generate X_V and Y_V are independent,

$$\begin{aligned} &P(X_B = z_B, Y_{V \setminus B} = z_{V \setminus B}) \\ &= \left(\sum_{\Phi_k^B} \rho_k \cdots \sum_{\Phi_i^B} (\rho_i + \eta \delta) \cdots \sum_{\Phi_1^B} \rho_1 \right) \left(\sum_{\tilde{\Phi}_k^{V \setminus B}} \tilde{\rho}_k \cdots \sum_{\tilde{\Phi}_j^{V \setminus B}} (\tilde{\rho}_j + \eta \tilde{\delta}) \cdots \sum_{\tilde{\Phi}_1^{V \setminus B}} \tilde{\rho}_1 \right) \\ &= (|\mathfrak{X}_B|^{-1} + \eta c_1 \lambda_{A_1} + O(\eta^2)) (|\mathfrak{X}_{V \setminus B}|^{-1} + \eta c_2 \lambda_{A_2} + O(\eta^2)) \\ &= |\mathfrak{X}_V|^{-1} + \eta (c'_1 \lambda_{A_1} + c'_2 \lambda_{A_2}) + O(\eta^2). \end{aligned}$$

It follows that

$$P(Z_V = z_V) = |\mathfrak{X}_V|^{-1} + \eta(c_1''\lambda_{A_1} + c_2''\lambda_{A_2}) + O(\eta^2),$$

for some $c_i'' > 0$. Then by an appropriate choice of scaling for each λ_{A_i} we see that $\Lambda_{A_1} + \Lambda_{A_2} \leq \text{TC}_0(\mathcal{G})$. For non-empty W , we can draw $Z_W = X_W = Y_W$ as a uniform random variable, and then look at $X_V | X_W$; the proof is otherwise the same. \square

Example 6.13. The bidirected 4-cycle in Figure 13(a) is not geared, and therefore we cannot apply our earlier results to it directly. The nested model for this graph, however, yields parametrizable sets

$$\mathcal{A}(\mathcal{G}) = \{1, 2, 12, 3, 23, 123, 4, 14, 124, 34, 134, 234, 1234\}$$

(these are just the bidirected-connected sets). The two subgraphs in Figures 13(b) and (c), say \mathcal{G}_1 and \mathcal{G}_2 , are geared, however, and have parametrizable sets

$$\mathcal{A}(\mathcal{G}_1) = \{1, 2, 12, 3, 23, 123, 4, 34, 234, 1234\}$$

$$\mathcal{A}(\mathcal{G}_2) = \{1, 2, 12, 3, 4, 14, 124, 34, 134, 1234\};$$

therefore $\bigoplus_{A \in \mathcal{A}(\mathcal{G}_i)} \Lambda_A \leq \text{TC}_0(\mathcal{G}_i)$ for $i = 1, 2$ by Corollary 6.10. Note that $\mathcal{A}(\mathcal{G}_1) \cup \mathcal{A}(\mathcal{G}_2) = \mathcal{A}(\mathcal{G})$, and therefore by applying Proposition 6.12 with these graphs, we find that

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A = \bigoplus_{A \in \mathcal{A}(\mathcal{G}_1)} \Lambda_A + \bigoplus_{A \in \mathcal{A}(\mathcal{G}_2)} \Lambda_A \leq \text{TC}_0(\mathcal{G}).$$

We are now in a position to prove the main result for general mDAGs.

Proof of Theorem 6.4. Suppose first that \mathcal{G} is geared.

$p[\rho_k, \dots, \rho_i + \eta\varepsilon_i, \dots, \rho_1]$ obeys the nested Markov property for any degenerate function ε_i and η sufficiently small that $1 + \eta\varepsilon_i$ is positive; it follows that $T_i \leq \text{TC}_0$ for each i , and that therefore using Corollary 6.10,

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq T_1 + \dots + T_k$$

is also contained in TC_0 , by the differentiability of $p[\cdot]$ at (ρ_k, \dots, ρ_1) .

Now for general \mathcal{G} , and each $A \in \mathcal{A}(\mathcal{G})$, there exists a geared subgraph \mathcal{G}' of \mathcal{G} such that $\Lambda_A \subseteq \text{TC}_0(\mathcal{G}')$ by Corollary 3.8. Then applying Proposition 6.12, we see that the space spanned by these subspaces is contained within the tangent cone for \mathcal{G} :

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq \text{TC}_0(\mathcal{G}).$$

If a distribution is in the marginal model then it is also in the nested model, and therefore TC_0 is contained within the tangent space TS_0^n of $\mathcal{N}(\mathcal{G})$ at p_0 , which has dimension

$$\begin{aligned}\dim(\text{TS}_0^n) &= \sum_{H \in \mathcal{H}(\mathcal{G})} (|\mathfrak{X}_H| - 1) \cdot |\mathfrak{X}_T| \\ &= \sum_{A \in \mathcal{A}(\mathcal{G})} \dim(\Lambda_A).\end{aligned}$$

Then combining

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq \text{TC}_0 \subseteq \text{TS}_0^n$$

with the dimension of TS_0^n gives the result. \square

7 Smoothness of the marginal model

The results of Section 6, together with the smoothness of the nested model, allows us to show that for geared graphs, the marginal model is smooth almost everywhere.

Theorem 7.1. *For a geared graph \mathcal{G} and state-space \mathfrak{X}_{VW} , the interior of the marginal model $\mathcal{M}(\mathcal{G})$ is a manifold of dimension $d(\mathcal{G}, \mathfrak{X}_{VW})$, and its boundary is described by a finite number of semi-algebraic constraints.*

Proof. The nested Markov model is parametrically defined, and therefore its Zariski closure is an irreducible variety (see, e.g. Cox et al., 2007, Proposition 4.5.5). Furthermore, there is a diffeomorphism between the set of strictly positive distributions obeying the nested Markov property, and an open parameter set. It follows that $\mathcal{N}(\mathcal{G})$ is a manifold on the interior of the simplex.

The marginal model is a semi-algebraic set, contained within the irreducible variety defined by the Zariski closure of the nested Markov model, so $\mathcal{M}(\mathcal{G})$ is a subset of $\mathcal{N}(\mathcal{G})$ defined by a finite number of additional polynomial inequalities. It follows that it is also a manifold at any point these inequality constraints are not active. \square

It follows from Theorem 7.1 that the interior of the marginal model for a geared mDAG is a curved exponential family of dimension $d(\mathcal{G}, \mathfrak{X}_{VW})$, and that therefore the nice statistical properties of these models can be applied. For example, the maximum likelihood estimator of a distribution within the model will be asymptotically normal and unbiased, and the likelihood ratio statistic for testing this model has an asymptotic $\chi^2_{|\mathfrak{X}_V| - d - 1}$ -distribution. For a point on the boundary defined by an active inequality constraint, the asymptotic distribution may be much more complicated (Drton, 2009b).

Inequality constraints are generally much more complicated than equality constraints, and efforts to characterize them fully have been limited by computational challenges. Evans (2012), generalizing a result first given by Pearl (1995), provides a graphical criterion for obtaining some inequalities, but deriving all such bounds may be an NP-hard problem (ver Steeg and Galstyan, 2011).

Geared mDAG models are semi-algebraic sets because they are given by variable elimination over a finite discrete latent variable model. However, for non-geared mDAGs we cannot assume that the latent variables are discrete without loss of generality, so it is conceivable that these marginal models may be defined by non-polynomial inequalities on the probabilities. We conjecture however, that a result akin to Theorem 7.1 does hold for general graphs.

7.1 Model Fitting

In theory we can fit the marginal model for a geared graph using a latent variable model of the kind derived in Section 5. In practice this model is massively over parameterized and unidentifiable, with the state-space of sets \mathcal{F}_v being potentially be very large even for modest graphs; this will cause problems for most standard fitting algorithms. However, for any graph \mathcal{G} —whether geared or not—and any latent variable model $\mathcal{L}(\mathcal{G})$, we have $\mathcal{L}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$. Fitting the nested model by maximum likelihood (ML) is straightforward using the algorithm in Evans and Richardson (2010), and many ML methods for fitting latent variable models are available. If the estimates for these two models agree, then we have found the maximum likelihood estimate (MLE) for the marginal model; if not, then we at least obtain a range of possible values for the log-likelihood at the MLE, and can use this to confirm or refute the marginal model.

References

- C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2007.
- D. Cox, J. Little, and D. O’Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2007. Third Edition.
- A. Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.
- A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.
- M. Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009a.

- M. Drton. Likelihood ratio tests and singularities. *Annals of Statistics*, 37(2):979–1012, 2009b.
- R. J. Evans. Graphical methods for inequality constraints in marginalized DAGs. In *Machine Learning for Signal Processing (MLSP)*, 2012.
- R. J. Evans. Graphs for margins of Bayesian networks. arXiv preprint: 1408.1809, 2014.
- R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th conference on Uncertainty in Artificial Intelligence (UAI-08)*, 2010.
- R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 42:1452–1482, 2014.
- R. J. Evans and T. S. Richardson. Parameterization of the discrete nested Markov model. In preparation, 2015.
- T. Fritz. Beyond Bell’s Theorem: Correlation scenarios. *New J. Phys*, 14:103001, 2012.
- S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.
- J. Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923. In Polish; English translation by D. Dabrowska and T. Speed in *Statist. Science* 5 463–472, 1990.
- J. Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the 11th conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 435–443, 1995.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009.
- T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 30(1):145–157, 2003.
- T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.
- T. S. Richardson, R. J. Evans, and J. M. Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

- I. Shpitser, T. S. Richardson, J. M. Robins, and R. J. Evans. Parameter and structure learning in nested Markov models. In *UAI-12 (Workshop on Causal Structure Learning)*, 2012.
- I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT press, 2000.
- G. ver Steeg and A. Galstyan. A sequence of relaxations constraining hidden variable models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*, 2011.
- T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 255–268, 1991.

A Technical Proofs

A.1 Proof of Proposition 4.2

Proof of Proposition 4.2. Suppose that $A, B \subseteq V$ are distinct sets, then we claim $\Lambda_A \perp \Lambda_B$; to see this, assume without loss of generality that there exists $a \in A \setminus B$, and take any $p \in \Lambda_A$ and $q \in \Lambda_B$.

Then using the fact that $q(x_V)$ does not depend upon x_a ,

$$\begin{aligned} \sum_{x_V} p(x_V) \cdot q(x_V) &= \sum_{x_{V \setminus a}} q(x_V) \sum_{x_a} p(x_V) \\ &= \sum_{x_{V \setminus a}} q(x_V) \cdot 0 \\ &= 0, \end{aligned}$$

so the claim holds.

Now, we claim that the vector space Λ_A has dimension at least $\prod_{a \in A} (|\mathfrak{X}_a| - 1)$. To see this, give each state-space \mathfrak{X}_a an element denoted 0, and let $\tilde{\mathfrak{X}}_a = \mathfrak{X}_a \setminus \{0\}$. We can freely pick values $p(x_A)$ for $x_A \in \tilde{\mathfrak{X}}_A$ as long as we then ensure that

$$p(x_{A \setminus a}, 0) = - \sum_{y_a} p(x_{A \setminus a}, y_a).$$

Lastly, note that counting up the dimensions of each Λ_A gives

$$\sum_{A \subseteq V} \prod_{a \in A} (|\mathfrak{X}_a| - 1) = \prod_{a \in V} |\mathfrak{X}_a|,$$

which is the same dimension as $\mathbb{R}^{|\mathfrak{X}_V|}$. Since the subspaces are all orthogonal, it follows that the direct sum gives the whole space. \square

A.2 Degenerate Functions

We present a series of Lemmas which build up to showing that we can construct degenerate functions from finite sums and products of degenerate functions with simpler argument sets.

Lemma A.1. *Let λ be a discrete $(A \cup B)$ -degenerate function, for $A \cap B = \emptyset$. Then λ can be written as a finite sum*

$$\lambda = \sum_i \lambda_A^i \lambda_B^i$$

of A -degenerate functions λ_A^i , and B -degenerate functions λ_B^i .

Proof. Since a matrix can be written as a sum of rank one matrices, clearly we can find (not necessarily degenerate) functions such that the result holds. But now suppose that the λ_A^i are not degenerate over $a \in A$, and consider

$$\begin{aligned} & \sum_i \left(\lambda_A^i(x_A) - \sum_{y_a} \lambda_A^i(x_{A \setminus a}, y_a) \right) \lambda_B^i(x_B) \\ &= \sum_i \lambda_A^i(x_A) \lambda_B^i(x_B) - \sum_{y_a} \sum_i \lambda_A^i(x_{A \setminus a}, y_a) \lambda_B^i(x_B) \\ &= \lambda(x_A, x_B) - \sum_{y_a} \lambda(y_a, x_{A \setminus a}, x_B) \\ &= \lambda(x_A, x_B). \end{aligned}$$

Thus we can replace each λ_A^i with the degenerate function

$$\tilde{\lambda}_A^i(x_A) \equiv \left(\lambda_A^i(x_A) - \sum_{y_a} \lambda_A^i(x_{A \setminus a}, y_a) \right)$$

and not affect the result. By repeating the argument we can assume that each λ_A^i is degenerate in every $a \in A$, and each λ_B^i degenerate in every $b \in B$. \square

Lemma A.2. *Let λ be a discrete $A \triangle B$ degenerate function. Then λ can be written as a finite sum*

$$\lambda = \sum_j \lambda_A^j \lambda_B^j$$

of A -degenerate functions λ_A^j , and B -degenerate functions λ_B^j .

Proof. Let $A' = A \setminus B$ and $B' = B \setminus A$ and $D = A \cap B$, so that $A \triangle B = A' \cup B'$, $A = A' \cup D$ and $B = B' \cup D$; note that A' , B' and D are all disjoint.

For each $y_D \in \mathfrak{X}_D$, define a degenerate function $\eta_D(\cdot; y_D) : \mathfrak{X}_D \rightarrow \mathbb{R}$ by

$$\eta_D(x_D; y_D) = \alpha^{-1} \prod_{d \in D} (|\mathfrak{X}_d| \mathbb{1}_{\{x_d=y_d\}} - 1).$$

where $\alpha = \prod_{d \in D} |\mathfrak{X}_d| \cdot (|\mathfrak{X}_d| - 1)$. One can verify easily that

$$\sum_{x_d \in \mathfrak{X}_d} \eta_D(x_D; y_D) = 0$$

for any y_D and $x_{D \setminus d}$, and that

$$\sum_{y_D \in \mathfrak{X}_D} \eta_D(x_D; y_D)^2 = 1;$$

in particular the last expression is independent of x_D .

Now, let λ be a discrete C -degenerate function, and using Lemma A.1 write it as

$$\lambda = \sum_{i=1}^j \lambda_{A'}^i \lambda_{B'}^i$$

where $\lambda_{A'}^i$ and $\lambda_{B'}^i$ are respectively A' and B' degenerate. Then for each $k \in \mathfrak{X}_D$, define $\lambda_A^{jk} = \lambda_{A'}^j \eta_D(\cdot; k)$ and $\lambda_B^{jk} = \lambda_{B'}^j \eta_D(\cdot; k)$. Clearly each of these is degenerate in $A = A' \cup D$ and $B = B' \cup D$ respectively. Further,

$$\begin{aligned} \sum_{i=1}^j \sum_{k \in \mathfrak{X}_D} \lambda_{A'}^{ik} \lambda_{B'}^{ik} &= \sum_{i=1}^j \sum_{k \in \mathfrak{X}_D} \lambda_{A'}^i \lambda_{B'}^i \eta_D(\cdot; k)^2 \\ &= \sum_{i=1}^j \lambda_{A'}^i \lambda_{B'}^i \sum_{k \in \mathfrak{X}_D} \eta_D(\cdot; k)^2 \\ &= \sum_{i=1}^j \lambda_{A'}^i \lambda_{B'}^i \\ &= \lambda. \end{aligned}$$

□

Lemma A.3. Let $\lambda : \mathfrak{X}_A \rightarrow \mathbb{R}$ be an A -degenerate function, and let $A = \bigtriangleup_{i \in I} A_i$ for some finite collection of sets $\{A_i : i \in I\}$. Then there exists a finite collection of A_i -degenerate functions $\lambda_i^j : \mathfrak{X}_{A_i} \rightarrow \mathbb{R}$ for $i \in I, j \in J$, such that

$$\lambda = \sum_{j \in J} \prod_{i \in I} \lambda_i^j.$$

Proof. This just follows from repeatedly applying Lemma A.2. □

A.3 Proof of Lemma 6.6

Lemma A.4. *Let \mathfrak{X} and \mathcal{Y} be finite sets, define $\mathcal{F} = \{f : \mathfrak{X} \rightarrow \mathcal{Y}\}$, and take $\lambda : \mathcal{Y} \rightarrow \mathbb{R}$. Then for any $A \subseteq \mathcal{Y}$ and $x \in \mathfrak{X}$,*

$$\sum_{\substack{f \in \mathcal{F} \\ f(x) \in A}} \lambda(f(x)) = |\mathcal{Y}|^{|\mathfrak{X}|-1} \sum_{y \in A} \lambda(y),$$

and if $x_1 \neq x_2$,

$$\sum_{\substack{f \in \mathcal{F} \\ f(x_1) \in A}} \lambda(f(x_2)) = |A| |\mathcal{Y}|^{|\mathfrak{X}|-2} \cdot \sum_{y \in \mathcal{Y}} \lambda(y).$$

In particular note that if λ is degenerate, the last expression is zero.

Proof. Clearly if $A = \mathcal{Y}$, then

$$\begin{aligned} \sum_{\substack{f \in \mathcal{F} \\ f(x) \in \mathcal{Y}}} \lambda(f(x)) &= \sum_{f \in \mathcal{F}} \lambda(f(x)) \\ &= |\mathcal{Y}|^{|\mathfrak{X}|-1} \sum_{y \in \mathcal{Y}} \lambda(y), \end{aligned}$$

since there are exactly $|\mathcal{Y}|^{|\mathfrak{X}|-1}$ functions in \mathcal{F} such that $f(x) = y$ for each $y \in \mathcal{Y}$. The first result follows in general by setting $\lambda'(y) = \lambda(y) \mathbb{1}_A(y)$.

The second result follows by similar combinatorial methods. \square

Proof of Lemma 6.6. It is clear that we only need prove the result for $E = \emptyset$, since we can just incorporate f_E as though they were observable parents of C , and the result is the same.

First consider the case $C = \{v\}$; let $L = \text{pa}_G(v)$ and take any set $K \subseteq L$. Let $\lambda : \mathfrak{X}_v \times \mathfrak{X}_K \rightarrow \mathbb{R}$ be a degenerate function, and for each $f : \mathfrak{X}_L \times \mathcal{F}_{\pi(x)} \rightarrow \mathfrak{X}_v$, define

$$\delta(f) = \sum_{\substack{y_L \in \mathfrak{X}_L \\ g_{\pi(v)} \in \mathcal{F}_{\pi(v)}}} \lambda(f(y_L, g_{\pi(v)}), y_K).$$

Then for fixed $x_v, x_L, f_{\pi(v)}$,

$$\begin{aligned} \sum_{\substack{f \in \mathcal{F}_v \\ f(x_L, f_{\pi(v)}) = x_v}} \delta(f) &= \sum_{\substack{f \in \mathcal{F}_v \\ f(x_L, f_{\pi(v)}) = x_v}} \sum_{\substack{y_L \in \mathfrak{X}_L \\ g_{\pi(v)} \in \mathcal{F}_{\pi(v)}}} \lambda(f(y_L, g_{\pi(v)}), y_K) \\ &= \sum_{\substack{y_L \in \mathfrak{X}_L \\ g_{\pi(v)} \in \mathcal{F}_{\pi(v)}}} \sum_{\substack{f \in \mathcal{F}_v \\ f(x_L, f_{\pi(v)}) = x_v}} \lambda(f(y_L, g_{\pi(v)}), y_K). \end{aligned}$$

But since λ is degenerate, the inner sum is zero unless both $x_L = y_L$ and $f_{\pi(v)} = g_{\pi(v)}$ by Lemma A.4. This leaves

$$\begin{aligned} &= \sum_{f(x_L)=x_v} \lambda(f(x_L), x_K) \\ &= |\mathfrak{X}_v|^{|\mathfrak{X}_L||\mathcal{F}_{\pi(v)}|-1} \cdot \lambda(x_v, x_K) \end{aligned}$$

where the constant represents the number of distinct functions $f \in \mathcal{F}_v$ such that $f(x_L, f_{\pi(v)}) = x_v$. Hence the result holds for $C = \{v\}$.

Now consider a general $C \subseteq R_i$; we prove the result by induction on the size of C . Given any $\text{sterile}_{\mathcal{G}}(C) \subseteq A \subseteq C \cup \text{pa}_{\mathcal{G}}(C)$, we first claim that we can write $A = A_1 \triangle A_2$ where for $\text{sterile}_{\mathcal{G}}(C_i) \subseteq A_i \subseteq C_i \cup \text{pa}_{\mathcal{G}}(C_i)$ for $i = 1, 2$ and disjoint non-empty C_1, C_2 with $C_1 \cup C_2 = C$.

To see this pick $C_2 = \{w\}$, $C_1 = C \setminus \{w\}$ for some $w \in \text{sterile}_{\mathcal{G}}(C)$, and then set $A_1 = (A \cup \text{sterile}_{\mathcal{G}}(C_1)) \cap (C_1 \cup \text{pa}_{\mathcal{G}}(C_1))$ and $A_2 = A \setminus A_1$. Clearly A_1 satisfies the required conditions. Since w was chosen to be sterile, $w \notin A_1$ and therefore $w \in A_2$; in addition, the only elements of A not contained in A_1 are those which are neither in C_1 nor $\text{pa}_{\mathcal{G}}(C_1)$; but since they are in $C \cup \text{pa}_{\mathcal{G}}(C)$, they must instead be in $\{w\} \cup \text{pa}_{\mathcal{G}}(w)$. Hence the claim holds.

Now first suppose that $\lambda = \lambda_1 \cdot \lambda_2$ for degenerate functions $\lambda_i : \mathfrak{X}_{A_i} \rightarrow \mathbb{R}$. By the induction hypothesis, we can find degenerate δ_1, δ_2 such that

$$\begin{aligned} \sum_{\substack{f_v(x,f)=x_v \\ v \in C_1}} \delta_1(f_{C_1}) &= c_1 \cdot \lambda_1(x_{A_1}) \\ \sum_{\substack{f_v(x,f)=x_v \\ v \in C_2}} \delta_2(f_{C_2}) &= c_2 \cdot \lambda_2(x_{A_2}). \end{aligned}$$

Then letting $E = R_i \setminus C$,

$$\begin{aligned} \sum_{\substack{f_v(x,f)=x_v \\ v \in R_i}} \delta_1(f_{C_1}) \cdot \delta_2(f_2) &= \sum_{\substack{f_v(x,f)=x_v \\ v \in E}} \sum_{\substack{f_v(x,f)=x_v \\ v \in C_1}} \sum_{\substack{f_v(x,f)=x_v \\ v \in C_2}} \delta_1(f_{C_1}) \cdot \delta_2(f_2) \\ &= c_0 \sum_{\substack{f_v(x,f)=x_v \\ v \in C_1}} \sum_{\substack{f_v(x,f)=x_v \\ v \in C_2}} \delta_1(f_{C_1}) \cdot \delta_2(f_2) \\ &= c_0 \sum_{\substack{f_{C_1} \in \mathcal{F}_v \\ f_{C_1}(\mathbf{x})=x_v}} \delta_1(f_{C_1}) \sum_{\substack{f_2 \in \mathcal{F}_w \\ f_2(\mathbf{x})=x_w}} \delta_2(f_2) \\ &= c_0 c_1 c_2 \cdot \lambda_1(x_{A_1}) \cdot \lambda_2(x_{A_2}). \end{aligned}$$

However a general degenerate function $\lambda : \mathfrak{X}_A \rightarrow \mathbb{R}$ can be written as a finite linear combination

$$\lambda = \sum_j \lambda_1^j \cdot \lambda_2^j$$

of degenerate functions $\lambda_i^j : \mathfrak{X}_{A_i} \rightarrow \mathbb{R}$, so the result follows by linearity of summations.

For the final part, note that if $v \in C \setminus B$, then the summation over Φ_i^B will include every function $f_v \in \mathcal{F}_v$. Then δ is degenerate and a function of f_v , so the sum is 0. On the other hand, if $v \in (R_i \cap B) \setminus C$, then δ is not a function of f_v and summing over all \mathcal{F}_v just involves $|\mathfrak{X}_v|$ extra identical terms in the summation. \square

A.4 Proof of Lemma 6.8

Proof of Lemma 6.8. First construct a directed graph Π^* on I_C in which $i \rightarrow j$ precisely when there exist $v_j \in R_j \cap C$ and $v_i \in R_i \cap C$ such that $v_j \in \pi(v_i)$. That Π^* is acyclic follows from the definition of π , which implicitly imposes a partial order on the R_j .

Let j be the maximal element of I_C ; we claim that for any other $i \in I_C$, there is always a directed path in Π_C from j to i . To see this, note that since C is bidirected-connected, there is a bidirected path in \mathcal{G} from some $v_j \in C \cap R_j$ to $v_i \in C \cap R_i$; given such a path, ρ , trim it so that only the end-points are in $C \cap R_j$ and $C \cap R_i$ respectively.

If ρ is just $v_j \leftrightarrow v_i$, then we are done, since $v_j \in \pi(v_i)$ by definition of π . Otherwise, ρ begins $v_i \leftrightarrow v_k \leftrightarrow \dots$ for some $v_k \in R_k \cap C$, where $i > k > j$. So we can apply an inductive argument to find a path from j to k in Π_C^* , and the edge $v_i \leftrightarrow v_k$ implies that $k \rightarrow i$ in Π_C^* .

Now, Π_C^* is a connected DAG with a unique root node j , so we can simply take any singly connected subgraph Π_C to fulfil the conditions of the lemma. \square